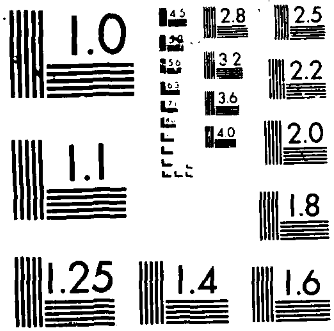
AD-A186 602 THE VALIDITY OF SELECTION AND CLASSIFICATION PROCEDURES 1/2
FOR PREDICTING JOB PERFORMANCE(U) INSTITUTE FOR DEFENSE
ANALYSES ALEXANDRIA VA J ZEIDNER APR 87 IDA-P-1977
UNCLASSIFIED MDA903-84-C-0031 F/G 5/9 NL


1.0
4.5
5.0
5.6
6.3
7.1
8.0
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

HQ-E500 907

2

IDA PAPER P-1977

# THE VALIDITY OF SELECTION AND CLASSIFICATION PROCEDURES FOR PREDICTING JOB PERFORMANCE

Joseph Zeidner

April 1987

*Prepared for*
Office of the Under Secretary of Defense for Research and Engineering

INSTITUTE FOR DEFENSE ANALYSES
1801 N. Beauregard Street, Alexandria, Virginia 22311

## DEFINITIONS

IDA publishes the following documents to report the results of its work.

### Reports

Reports are the most authoritative and most carefully considered products IDA publishes. They normally embody results of major projects which (a) have a direct bearing on decisions affecting major programs, or (b) address issues of significant concern to the Executive Branch, the Congress and/or the public, or (c) address issues that have significant economic implications. IDA Reports are reviewed by outside panels of experts to ensure their high quality and relevance to the problems studied, and they are released by the President of IDA.

### Papers

Papers normally address relatively restricted technical or policy issues. They communicate the results of special analyses, interim reports or phases of a task, ad hoc or quick reaction work. Papers are reviewed to ensure that they meet standards similar to those expected of refereed papers in professional journals.

### Memorandum Reports

IDA Memorandum Reports are used for the convenience of the sponsors or the analysts to record substantive work done in quick reaction studies and major interactive technical support activities; to make available preliminary and tentative results of analyses or of working group and panel activities; to forward information that is essentially unanalyzed and unevaluated; or to make a record of conferences, meetings, or briefings, or of data developed in the course of an investigation. Review of Memorandum Reports is suited to their content and intended use.

The results of IDA work are also conveyed by briefings and informal memoranda to sponsors and others designated by the sponsors, when appropriate.

The work in this document was conducted under contract MDA 903 84 C 0031 for the Department of Defense. The publication of this IDA Paper does not indicate endorsement by the Department of Defense, nor should the contents be construed as reflecting the official position of that agency.

This paper has been reviewed by IDA to assure that it meets high standards of thoroughness, objectivity, and sound analytical methodology and that the conclusions stem from the methodology.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

## REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY
NA

2b. DECLASSIFICATION/DOWNGRADING SCHEDULE
NA

3. DISTRIBUTION/AVAILABILITY OF REPORT
Approved for public release; distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
IDA Paper P-1977

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
Institute for Defense Analyses

6b. OFFICE SYMBOL (If applicable)

7a. NAME OF MONITORING ORGANIZATION
DoD-IDA Management Office, OUSDRE

6c. ADDRESS (City, State, and Zip Code)
1801 N. Beauregard Street
Alexandria, VA 22311

7b. ADDRESS (CITY, STATE, AND ZIP CODE)
1801 N. Beauregard Street
Alexandria, VA 22311

8a. NAME OF FUNDING/SPONSORING ORGANIZATION
OUSDRE/R&AT

8b. OFFICE SYMBOL (If applicable)

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
MDA 903 84 C 0031

8c. ADDRESS (City, State, and Zip Code)
Room 3D129
The Pentagon
Washington, DC 20301-3080

10. SOURCE OF FUNDING NUMBERS

| PROGRAM ELEMENT | PROJECT NO. | TASK NO. | WORK UNIT ACCESSION NO. |
|---|---|---|---|
|  |  | T-D2-435 |  |

11. TITLE (Include Security Classification)
The Validity of Selection and Classification Procedures for Predicting Job Performance

12. PERSONAL AUTHOR(S).
Joseph Zeidner

13. TYPE OF REPORT
Final

13b. TIME COVERED
FROM 8-86 TO 4-87

14. DATE OF REPORT (Year, Month, Day)
April 1987

15. PAGE COUNT
136

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

| FIELD | GROUP | SUB-GROUP |
|---|---|---|
|  |  |  |
|  |  |  |

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
selection, classification, job performance, test validity, selection validity, classification validity

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This report reviews major validation studies and meta-analytical summaries to assess the effectiveness of selection and classification procedures for predicting job performance in military and civilian settings. Initially, the average job validity coefficient across all jobs was computed to be in the low .20s. Currently, when a more uniform collection of studies is considered, statistical artifacts have been corrected, and carefully developed job criteria are used, the average job validity increased to the low .60s. Findings support major validity generalization concepts; however, job complexity and also criterion dimensions within a job both moderate validity. Expanding the predictor space of ASVAB and criterion dimensions appears to offer promise of differential validity for assignment.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
□ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. □ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
UNCLASSIFIED

22a. NAME OF RESPONSIBLE INDIVIDUAL
Jesse Orlansky

22b. TELEPHONE (Include Area Code)
(703) 578-2836

22c. OFFICE SYMBOL

83 APR edition may be used until exhausted.
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

IDA PAPER P-1977

# THE VALIDITY OF SELECTION AND CLASSIFICATION PROCEDURES FOR PREDICTING JOB PERFORMANCE

Joseph Zeidner

April 1987

INSTITUTE FOR DEFENSE ANALYSES
Contract MDA 903 84 C 0031
Task T-D2-435

# ABSTRACT

This report reviews major validation studies and meta-analytic summaries to assess the effectiveness of selection and classification procedures for predicting job performance in military and civilian settings. Initially, the average job validity coefficient across all jobs was computed to be in the low .20s. Currently, when a more uniform collection of studies is considered, statistical artifacts have been corrected, and carefully developed job criteria are used, the average job validity increases to the low .60s. Findings support major validity generalization concepts; however, job complexity and also criterion dimensions within a job both moderate validity. Expanding the predictor space of ASVAB and criterion dimensions appears to offer promise of differential validity for assignment.

# ACKNOWLEDGMENTS

I wish to express my appreciation to Jesse Orlansky, Institute for Defense Analyses, for suggesting that the validity and utility of selection and classification procedures be evaluated, for providing me with valuable information, and for his useful suggestions on preparing this report.

The help uf Arthur Drucker, Army Research Institute, and Dexter Fletcher, Institute for Defense Analyses, in editing is gratefully acknowledged.

John Campbell, University of Minnesota, Newell Eaton, Army Research Institute, Milton Maier, Center for Naval Analyses, and Frank Schmidt, University of Iowa, provided many useful technical comments for which I am grateful.

I am most appreciative of the many discussions over the years with Cecil Johnson, formerly at the Army Research Institute, as we struggled over classification issues, especially potential allocation efficiency, and for his comments on the draft of the manuscript.

The responsibility for any remaining errors is, of course, mine.

This study was performed for the Office of the Under Secretary of Defense (Acquisition)/(Research and Advanced Technology). Technical cognizance for this work under Task Order T-D2-435, was assigned to the Assistant for Training and Personnel Technology. We are indebted to Captain Paul R. Chatelier, USN, now retired, and to Earl A. Alluisi for their contributions to and interest in this effort.

## CONTENTS

## TABLES

FIGURE

# ABBREVIATIONS

| | |
|---|---|
| ACB | Airmen Classification Battery |
| AGCT | Army General Classification Test |
| AI | Aptitude Index |
| AR | Arithmetic Reasoning |
| AS | Auto Shop Information |
| ASVAB | Armed Services Vocational Aptitude Battery |
| B&C | Business and Clerical |
| CL | Clerical/Administrative |
| CO | Combat |
| CS | Coding Speed |
| E&E | Electronic and Electrical |
| EI | Electronic Information |
| EL | Electronic Repair |
| F | Finger Dexterity |
| FA | Field Artillery |
| G | General Intelligence |
| GATB | General Aptitude Test Battery |
| GM | General Maintenance |
| GS | General Science |
| GVN | Cognitive Ability |
| HS&T | Health, Social and Technology |
| K | Motor Coordination |
| KFM | Psychomotor Ability |
| M | Manual |
| M&C | Mechanical and Craft |
| MC | Mechanical Comprehension |
| MK | Mathematical Knowledge |
| MM | Mechanical |

| | |
|---|---|
| MOS | Military Occupational Specialties |
| N | Numerical Aptitude |
| NO | Numerical Operations |
| OF | Operations/Food |
| P | Form Perception |
| PAE | Potential Allocation Efficiency |
| PC | Paragraph Comprehension |
| Q | Clerical Perception |
| S | Spatial Aptitude |
| SC | Surveillance/Communications |
| SQT | Skill Qualification Test |
| SRQ | Perceptual Ability |
| ST | Skilled Technical |
| V | Verbal Aptitude |
| VE | General Verbal Ability |
| WK | Word Knowledge |

# SUMMARY

Analyses of major validation studies over the last half century along with recent meta-analytic reviews indicate that the magnitude of operational or true validity of selection tests has been systematically underestimated and that validity findings have been distorted by conceptual and methodological limitations.

It was traditionally believed that the criterion-related validity of a selection test was specific to a given situation of a job and that, therefore, an empirical validation was required for each new application. Concepts on how to generalize validity and selection procedures began to evolve during the 1970s. When statistical artifacts were taken into account, standardized cognitive ability tests were found to be valid predictors of performance for all jobs; it was also found that test validities were not specific to variations in job content or organizational context.

New methods of cumulating findings across studies were developed as well as methods of correcting variance across studies for sampling error and of correcting both the mean and variance for unreliability and range restriction. Numerous empirical studies during the last decade showed that differences in validity across studies were largely artifactual. Validity generalization provided a new framework for evaluating research findings and for understanding the predictive power of selection procedures.

Ghiselli summarized job performance validity data from the 1920s through the early 1970s. He found that the grand average

for all jobs taken together produced a validity coefficient in the low .20s. These validities were for individual tests, uncorrected for statistical artifacts. Hunter reanalyzed these same data, corrected for criterion unreliability and range restriction, and combined tests in weighted composites. The average validity coefficient increased to the high .40s.

Using a more uniform collection of studies that contain analyses of thousands of validity coefficients for the U.S. Employment Service's General Aptitude Test Battery and for the military's Armed Services Vocational Aptitude Battery (ASVAB) against job performance criteria, we found an average validity coefficient in the mid .50s.

When the ASVAB was validated against very carefully defined and measured job criteria designed to minimize the usual problems of reliability, criterion deficiency and contamination (measuring too little or too much) of existing performance measures, an average validity coefficient in the low .60s was found.

Combining ASVAB general cognitive ability subtests with alternative predictors, a composite with an average validity coefficient in the mid .60s was obtained against the same carefully developed criteria.

In less than a decade, empirical data have clearly confirmed the power of selection and classification procedures for predicting job performance and thus for increasing the productivity of the work force. Taken as a whole the present analysis supports the view of cognitive ability tests as being the best overall predictors of performance for entry level job performance. Validity can be increased by combining data from different types of tests, e.g., psychomotor, perceptual biodata, and temperament measures with general ability tests in a weighted composite.

The current ASVAB does not possess differential validity or potential allocation efficiency (PAE) as a means of increasing average job performance of assigned manpower. Recent results of Project A (Army Research Institute) indicate that PAE may be possible by considering simultaneously both expanded predictor and criterion domains and alternative ways of making use of PAE in a revised ASVAB should be investigated.

While the major conclusions of validity generalization appear sound, e.g. detailed job analyses are not required to place jobs into job families to find valid job predictors, both job complexity (as measured by information processing demands of the job) and the performance constructs defining a job, e.g. supervisory ratings, hands-on tests, and job knowledge tests moderate test validities. Thus current findings on validity generalization cannot support the view that validities of cognitive tests are the same across jobs varying in complexity or that test validities are the same against various performance measures of a job.

Using valid selection and classification procedures in the acquisition of large numbers of personnel results in very sizable productivity gains measured in dollar terms. For example, in 1973, the yearly gain attributable to employing highly valid selection and classification ability tests in recruiting military enlisted personnel was $442 million; and in 1986, the yearly potential impact of employing highly valid cognitive ability tests in hiring new federal employees was $8 billion.

# I. INTRODUCTION

This report is the first of two planned reports evaluating the utility of standardized testing. This report (the first) is concerned with the validity of selection and classification procedures for predicting job performance in military and civilian settings; the second report considers the economic benefits of predicting job performance. (See Appendix A for a glossary of terms.)

The present report reviews major validation studies and meta-analytic summaries in order to: assess the effectiveness of tests and alternative predictors; provide central tendencies of test-criterion combinations along with the range of validities expected for different uses; and suggest ways in which traditional validation methodologies and concepts need to be modified.

We start with a brief description of some recent developments which have significantly contributed to our understanding of theory and practice in selection and classification.

## A. EARLY DEVELOPMENTS

One of the most enduring effects of the World War I personnel selection program must be the impetus it gave to mental testing. The Army Alpha tests were the first written tests of mental ability to gain respect and they still serve as the model for scientific testing today. Because tests were administered to groups, they represented a convenient means of ranking everyone for nearly every purpose. Employers were quick to utilize tests as one means of increasing productivity, especially since the tests were perceived as being objective and predictive of later performance. Over the decades numerous validation studies attested to their effectiveness as predictors of training and job success.

## B. SOCIAL CONCERNS

However, in recent decades there has been much social and scientific controversy surrounding testing. Critics have focused on tests' fairness and their adverse impact on examinees, their limited predictive powers for long-term job performance and the often narrow range of skills covered by them. In short, tests are criticized as inadequate for many of the purposes they were designed to serve. At the same time, scientific critics began to question the theoretical bases of measuring individual differences in cognitive skills, the inability of researchers to break the asymptomatic barrier of job validities (the ".3 problem"), and the limited advancements in theory and practice.

There has been a continuing concern that selection tests deny qualified applicants access to jobs. Title VII of the Civil Rights Act of 1964 has been the primary legal basis for protecting individuals against employment discrimination. The Tower amendment to the act, however, expressly permits the use of professionally developed ability tests in selecting employees. The Supreme Court laid down a series of rulings on test usage that together with the Equal Employment Opportunity Commission Uniform Guidelines define acceptable practices, particularly for demonstrating job relatedness and equal effectiveness in prediction for minorities and nonminorities.

The legal challenge to testing stimulated an interest in evaluating differential prediction in academia, industry, and the military through comparison of regression systems for different groups.

## C. VALIDITY GENERALIZATION

Case law also awakened a long-dormant interest in validity generalization or transportability of tests. The prevailing view through the years has been that employment test validations were situation-specific and that empirical data were needed for each new situation. Recent work correcting for various sources of artifactual, between-study variance, sought to support the utility of validity generalization and thus make it possible to develop general principles for linking ability tests to classes of jobs.

Schmitt and Schneider (1983) in their view of issues concerning validity generalization comment:

> Certainly, the research of Schmidt, Hunter and their colleagues has had more impact on personnel psychology than any other research reported. It has the potential for producing major differences in the way industrial/organizational psychologists approach a variety of problems, as well as providing a substantial scientific base for individual differences in job performance and ability. Their work has been instrumental in rethinking nearly every part of what is viewed as the traditional test validation model outlined previously...
>
> ...but we feel it is important that some issues concerning the Schmidt-Hunter research be raised in the hope that additional research, and/or careful review of existing information, will provide more substantial support for some of the Schmidt-Hunter assertions, or at least indicate appropriate caution, (pp. 108-109).

## D. CRITERION ISSUES

Most often tests and alternative selection procedures were validated against available criterion measures, with little effort made to evaluate the criterion measure itself. Using available criteria almost always resulted in measures that: were deficient (perhaps appropriate and relevant, but incomplete, e.g., using training grades in place of job performance measures); or were contaminated (perhaps included too much, e.g., using global ratings, with their emphasis on interpersonal factors, as an index of job proficiency).

\ serious shortcoming of most criterion measures is their low reliabilities resulting in a large downward bias in validity coefficients. In criterion-related validity studies, ratings traditionally have been the most frequently used measure of performance, with inter-rater agreement, according to Schmidt, Hunter, and Outerbridge (1986), averaging below r = .60. Corrections for criterion unreliability that better reflect true validities are generally not made.

The issue of validity, of course, is central for selection and classification since it tries to answer the question of what is being measured and how well it is being measured. The APA Standards for Educational and Psychological Tests (1985) addresses the significance of criterion-related validity measures:

> Criterion-related evidence demonstrates that test scores are systematically related to one or more outcome criteria. In this context the criterion is the variable of primary interest ... The choice of criterion and the measurement procedures used to obtain criterion scores are of central importance. Logically, the value of a criterion-related study depends on the relevance of the criterion measure that is used, (p. 11).

For the most part, however, there was little attention paid or support given to the criterion measure even though it was well recognized that criterion-related research depended on the quality of the criterion measure employed. In the military, from the end of World War II until well into the decade of the 1970s the criteria employed primarily consisted of administrative information from the files, training grades, and various types of peer and supervisory ratings.

By the mid 1970s the situation began to change. The longstanding aspiration of researchers to use something more comprehensive and relevant than training indicators or ratings of job performance as criteria for evaluating selections tests began to receive support. Again, in part, because of the legal emphasis on empirical measure of test validity against job performance, decisionmakers and scientists urned their attention to the difficult, time consuming, and expensive task of measuring job performance through a combination of objective hands-on measures of performance, job knowledge measures, and behaviorally anchored rating scales (BARS). Research results on BARS, however, show them to be no better psychometrically than other rating methods (Dunnette and Borman, 1979; Jacobs, Kafry, and Zedeck, 1980; and Schwab, Heneman, and De Cotiis, 1975).

On the technical side, research that focused on the development of predictors and job performance criteria had to provide not only greater understanding but improved predictive power of standardized selection procedures. At the same time, it was also recognized, with regard to social policy as Haney (1982) puts it:

> ....that while the role of standardized testing often is both advocated and challenged in technical terms, the prominent social concerns surrounding standardized testing, both now and in the past, are rooted in matters of social and political values, (p. 1032).

## E. COMPUTERIZED ADAPTIVE TESTING

Although research on "tailored testing" started several decades ago, the everyday application of computerized adaptive testing (CAT) only became possible with advances in microcomputer technology and refinement in Item Response Theory. CAT permits automated testing using a display screen and a light pen (and other devices) for responding. Test questions are tailored by the response to the previous question and computer-scored after each response. The terminal used by the examinee is designed expressly for testing purposes. The sequencing of items in tailored testing has as its principal goal equal precision of estimating ability for the total distribution of examinees, not just at the middle or at a given cut-score. Other CAT advantages are test security, simplicity of test revision, scoring accuracy, improved test reliability, reduced test administration costs and efficient use of time. Another significant potential of CAT is that it provides the capability for the use of entirely new types of tests via computer displays and input-output devices. Department of Defense has an ongoing large-scale implementation program designed to replace traditional paper-and-pencil tests with CAT.

## F. COGNITIVE ASSESSMENT

Until recent times, the theory of cognitive abilities in differential psychology depended on factor-analytic techniques. Thurstone's primary mental ability structure (or variations of it), with its seven relatively independent factors, served as the theoretical basis for selection and classification batteries for a half-century. Many cognitive psychologists, however, were looking for a deeper understanding of individual differences in information processing based on an experimental rather than on a correlational approach. In the cognitive approach, response to stimulus variation within an individual is examined more closely than variations of individuals to a given stimulus, which is the focus of the differential approach.

Research is now underway to see to what extent the two approaches can form a common basis for testing abilities for selection and training. The hope is that psychometric testing can be supplemented by information processing procedures - that there will be a convergence of the "two disciplines of scientific psychology." While computer technology now makes this possible, the critical question that remains to be demonstrated is whether there is improved validity.

## G. UTILITY

The idea of determining the utility of testing in cost-effectiveness terms is not new. Brogden (1949) demonstrated how the selection-ratio and the standard deviation of job performance in dollar terms can affect the economic benefit of selection tests. Brogden and Taylor (1950) stated that the criterion should measure the contribution of the individual to organizational productivity rather than the individual's contribution in terms of latent skills. Schmidt, Hunter, McKenzie, and Muldrow (1979) developed practical procedures for obtaining rational estimates of the standard deviation of performance in dollar terms and Schmidt, Hunter, and Pearlman (1982) provided evidence to support their rational approach. This work has stimulated a great deal of behavioral research in developing new or improved methods of utilizing selection and classification strategies and also in applying costing to other human resource areas.

The full utilization of both selection and classification data, as needed in the military, requires a person-job matching system to inventory available abilities for jobs and to develop a strategy for allocating those abilities to meet organizational goals. The development of a new job matching system would use performance criteria and predictor information, models for planning, executing and evaluating person-job decisions, and decision support systems such as data bases, communication interfaces, and control modules to achieve management objectives.

## H. PERFORMANCE MEASUREMENT IN THE MILITARY

The behavioral research laboratories within the military are giving considerable attention to [illegible] the issues just described. These issues are addressed [illegible] analysis of a common selection and classification battery, the Armed Services Vocational Aptitude Battery (ASVAB), which is administered to about one million military applicants each year in all four services. A version of the ASVAB also is given to about the same number of high school students for vocational counseling and recruiting purposes. Although the ASVAB is a direct lineal descendant of the Army Alpha of 1917, the services are now more active than ever before in seeking improvements in the Battery. Several influences are at work, including significant developments in cognitive theories and computer technologies, congressional directives that the ASVAB be shown to be valid against job performance (rather than training performance), and social concerns surrounding testing.

The behavioral research laboratories in the military are now undertaking cooperative research to develop improved job performance measures of enlisted personnel. A Committee on the Performance of Military Personnel of the National Research Council has recently issued a report on the status of the program (Wigdor and Green editors, 1986). They write:

> The Joint-Service Job Performance Measurement/ Enlistment Standards Project represents a landmark in the measurement of human performance. It is a demonstration project of unmatched scale and breadth of coverage. Although there have been potentially damaging problems with funding, the financial resources committed to the project far exceed what is available in the private sector. Perhaps more to the point, the Joint-Service Project is bringing criterion research a degree of systematic, scientific attention that it has too seldom received. Technically, the project is in the process of transforming the pivotal issue in criterion research from that of demonstrating the validity of a particular measure to the more complex task of comparing the substantive and psychometric adequacy of alternative criterion measures, (p. 1).

Dunnette and Borman (1979) in discussing the need for criterion research also note that while a selection system validation requires good performance measures, "The criterion has been with us forever and has received much attention," (p. 486). One continuing theme they identify as promising is to base measurement on conceptual and methodological guidelines of construct validation to avoid incompleteness and spuriousness. As Tenopyr and Oeltjen (1982) note, however, supervisory ratings continue to be the most common criterion against which tests are evaluated, with much emphasis on rating formats and little on the rating context. Schmitt and Schneider (1983) concluded that the "criterion problem" has certainly not been solved, but that "the most hopeful sign has been recent efforts to conceptualize, and conduct research on, criteria from a more nomological vantage point," (p. 100). Such an approach may involve cognitive models of the rating process, and conceptualizations of individual behavior over time, and thoughts about organizational participation.

The above themes stand in sharp contrast to the current approach being taken by the Joint-Service Job Proficiency Measurement/Enlisted Standards Project. The National Research Council Committee states:

> At the project's conceptual core is the assumption that the most direct measure of job performance is also the most valid. This assumption has led to a project preference for measuring "manifest, observable job behaviors" as opposed to less direct indicators, such as training grades, or less tangible characteristics, such as motivation or underlying abilities. Furthermore, the behavior of interest is "proficiency," what the Services call the "can-do" component of job performance, and not the entire array of possible behaviors that determine whether a person does do the job, (p. 15).

Hands-on tests in this context function as a benchmark or standard since they are the most faithful representation of actual job performance. The Army, however, (see Project A description later in this report) considers hands-on tests as just one of a number of types of measures that load on different factor dimensions, e.g., specific job skills, general soldiering, leadership, and effort.

The Committee feels that the largely psychometric emphasis on individual difference criterion measurement for validation purposes should be augmented by an absolute assessment of job competence or job mastery. Such an assessment should show how much of the whole job an individual can do and hopefully would lead to less misunderstanding by policy makers. The Marine Corps has research underway which attempts to combine such norm-referenced and criterion-referenced approaches.

## I. HISTORICAL MILESTONES

In closing this section on significant recent contributions it is of interest to note Dunnette and Borman's (1979), pp. 478-482, list of what they believe to be the most important milestones over the last 60 years (until 1977). I agree that their milestones are very significant developments.

For convenience the Dunnette and Borman list has been adapted and subdivided into three groups:

1. Technical Milestones

- Early Developments in Selection Research Technology (Texts published in the 1920s specified selection research procedures.)

- Individual Diagnosis and Vocational Counseling (Differential psychology principles and measures form the basis of individual job match.)

- Factor Analysis of Human Attributes (Methodologies that contributed to taxonomies of skills.)

- Selection Technology in World War II (New techniques for validation, job analysis, job performance and statistics were developed.)

- Critical Incidents Method (Critical incidents methodology defined jobs behaviorally.)

- Standards for Developing and Evaluating Tests (Standards, first published in 1954, classified concepts and provided validation guidelines.)

- Nonlinear Prediction Models in Selection Research. (Increased research on nonlinear prediction systems.)

- Simulations and Multiple Assessment Procedures (Extensive and expensive techniques for evaluation and development of personnel.)

2. Utilization Milestones

   - Large Scale Programs of Industrial Selection Research (Selection research programs initiated at first by a few major corporations.)

   - Growth of Test Publishing Industry (Test publishing has become big business with over 3000 measures.)

   - Decision Theory in Selection Research (Utility in terms of costs and benefits for selection strategies.)

3. Social Policy Milestones

   - Growing Political Emphasis on Equality of Opportunity (Legal efforts starting in the 1950s against employment discrimination.)

   - Testing, Selection Research and Civil Rights (Title VII of Civil Rights Act 1964 made selection research a matter of public and legal concern.)

   - Affirmative Action Programs by Employers (Personnel selection and classification procedures can assist in the best possible utilization and conservation of human resources.)

## II. REVIEW OF MAJOR VALIDATION STUDIES

Over the past 65 years, an enormous amount of data has been published on test validity. Because of the sheer volume of data reported, investigators have felt it of value to publish summaries of validity information for this reason alone. Edwin Ghiselli made it a life goal to develop taxonomies of validity data in as simple and condensed form as possible, and by integrating validity results he hoped to develop principles of personnel selection.

In recent years, with the development of meta-analytic procedures, several new reviews have emerged, analyzing the validities of various sub-groupings. As a consequence, many of the traditional concepts in selection have been challenged; new theoretical positions have been vigorously argued. The team of John Hunter and Frank Schmidt and colleagues has been preeminent in carrying out these new efforts which have provided strong evidence for the feasibility of making validity generalizations--the ascribing of similar test validities across different situations within broad job families.

This report presents a review of major criterion-related studies and meta-analytic validation summaries from the 1920s to date. It was done for a number of inter-related objectives:

a. To assess the effectiveness of tests and other predictors for use in personnel selection and classification,

b. Since all cost-effectiveness formulations of utility or test value must use an observed or estimated validity correlation coefficient, another purpose was to suggest the range of validities to be expected in different applications,

c. To catalogue central tendencies of validity coefficients of various test types against likely criteria for relevant occupational groups,

d. To provide practical bases for examining ways in which traditional test validation concepts need to be modified.

This review will include a much larger representation of published military validation studies than is to be found in typical journal reviews and furthermore will contrast results of the newer meta-analytic procedures with those of traditional (older) procedures.

Table 1 lists twelve major validation summary studies that are reviewed in this report, military and civilian. Although the summaries are evaluated independently, when considered together they provide fairly stable estimates of the predictive power of tests and alternative selection procedures.*

Numerous additional reviews and meta-analytic summaries of test-criterion combinations across jobs could have been included in this review; for example, Asher and Sciarrino (1974), Cohen, Moses, and Byham (1974), Cunnette (1972), Kane and Lawler (1978), Lilenthal and Pearlman (1983), O'Leary (1980), Pearlman (1982), Pearlman, Schmidt, and Hunter (1980), Schmidt, Gast-Rosenberg, and Hunter (1980), and Schmidt, Hunter, and Caplin (1981). Studies omitted were either included in other meta-analytic

*It should be noted that the work of John Hunter and/or colleagues has been directly included four times in this review. Additionally, several meta-analytic studies included in Hunter's reports have also been included in this report, but in greater detail. Without Hunter's and Frank Schmidt's empirical analyses, no contemporary reviewer of the military or civilian validation literature could have as many integrated, comprehensive results on hand, nor as much understanding of the meaning of the results.

TABLE 1. REVIEW OF MAJOR VALIDATION STUDIES

| Context | Description | Source |
|---|---|---|
| Military Studies | | |
| 1. Army General Classification Test (AGCT) (1940-1945) | Uses World War II samples in evaluating tests for training | PRS, ARMY 1945 |
| 2. Armed Services Vocational Aptitude Battery (ASVAB) (1980-1983) | Employs a very large sample in unique ongoing Army effort and uses professionally developed job-performance measures | McLaughlin, et al. 1984 |
| 3. Airman Classification Batteries (ACB) (1948-1975) | Evaluates the Air Force selection and classification battery for training | Weeks, et al. 1975 |
| 4. ASVAB Validation Across Services (1980-1985) | Evaluates ASVAB validity and differential validity in all services | Hunter, et al. 1985 |
| 5. Prediction of Military Job Performance (1952-1980) | Cumulates job performance results in 114 published military studies | Vineberg and Joyner 1982 |
| Civilian Studies | | |
| 6. Summary of Published and Unpublished Validities (1920-1971) | Categorizes thousands of validity studies by 20 types of tests, 21 job families, and two criteria | Ghiselli 1973 |
| 7. Reanalysis of Ghiselli's Results (1920-1971) | Meta-analysis of validities by job families arranged in order of decreasing cognitive complexity | Hunter 1981 |
| 8. General Aptitude Test Battery (1938-1983) | Compares observed versus true validities of the U.S. Employment Service's battery across job families | Hunter 1983 |

(Continued)

TABLE I. REVIEW OF MAJOR VALIDATION STUDIES (Continued)

| Context | Description | Source |
|---|---|---|
| Alternative Predictors | | |
| 9. Alternative Selection Procedures (1970-1979) | Reviews eight categories of selection procedures other than conventional measures of ability | Reilly and Chao 1982 |
| 10. Meta-analyses of Validity Studies (1964-1982) | Analysis of predictors criteria, predictor-criterion combinations, and estimates of variances | Schmitt, et al. 1984 |
| 11. Meta-analytic Comparisons of Predictors of Job Performance (1973-1984) | Compares previous and new meta-analytic studies of validity for major types of predictors and criteria | Hunter and Hunter 1984 |
| 12. ASVAB Validation using Multiple Job Criteria (1984-1987) | Estimates differential prediction across jobs for major domains of predictors and job performance | Campbell and ARI 1987 |

reviews described in this report, covered a more restricted sample of jobs or were largely confirmatory of validity generalizations. A few studies were included in this review for more detailed discussion and in some meta-analytic studies which were also summarized. The intent was to provide a broadly representative sample of significant military and civilian results of validity magnitudes obtained and comparisons among alternative predictors for various domains of job performance.

## A. ARMY GENERAL CLASSIFICATION TEST

As a matter of historic interest, this review begins with the Army General Classification Test (AGCT) results of World War II. The AGCT, an index of learned cognitive ability, was first used by the Army in October 1940 to facilitate assignment to training and jobs. The ready acceptance of tests in the military community preceding and during the Second War could be attributed to their successful use during World War I (Army Alpha and Beta tests) and to the broad acceptance of tests in industry and academia during the 1920s and 1930s. More than nine million individuals took one form or another of the AGCT by war's end. It was later released to the public for civilian use and as late as the 1970s some foreign military services were requesting and using it.

The AGCT consisted of 150 multiple-choice vocabulary, arithmetic, and block counting items. Standard scores with a mean of 100 and standard deviation of 20, were constructed from raw scores, and then distributed into five Army mental grades (see Table 2). Revisions in which part scores were recorded for the first time were offered in April 1945.

These revisions contained four subtests--reading and vocabulary, arithmetic computation, arithmetic reasoning, and pattern analysis.

TABLE 2. ARMY MENTAL GRADE GROUPS AND STANDARD SCORE RANGES

| Army Grade (Mental Group) | Standard Score Range | AFQT Percentile Score Range[a] |
|---|---|---|
| I | 130 and higher | 93-99 |
| II | 110 - 129 | 65-92 |
| III | 90 - 109 | 31-64 |
| IV | 70 - 89 | 10-30 |
| V | 69 and lower | 9 and lower |

Note. Percentage of the Army population falling into each mental group varied from time to time with changes in norms. In July 1942, the Group IV lower limit was changed from 70 to 60 to correspond better with the distribution anticipated from operational use. This grading system has remained with the Army to the present.

[a] The Armed Forces Qualification Test (AFQT) is currently an operational aptitude composite used to select enlistees for all services. It consists of subtests of the Armed Services Vocational Aptitude Battery: Work Knowledge, Paragraph Comprehension, Arithmetic Reasoning and Numerical Operations. In general, current policy is to accept only applicants who achieve a mental category III percentile score of 31 or higher for service.

AGCT was quite successful in selecting men for specialist training as evidenced by the magnitude of the many hundreds of validity coefficients obtained, a few of which are shown in Table 3 (PRS Staff, 1945). Since most of the samples for these studies had been preselected on the AGCT or on some highly correlated factor, the obtained relationships were, in general, quite restricted (see means and SDs in Table 3) and thus considerably underestimated the operational or true effectiveness of the tests. Data such as that in Table 3 strongly contributed to the concept of situational specificity, i.e., validities of the same tests for the same jobs but in different settings varied because of subtle differences in job requirements.

TABLE 3. VARIOUS EXAMPLES OF VALIDITY COEFFICIENTS FOR AGCT

| Population | Criterion | N | Mean | SD | r |
|---|---|---|---|---|---|
| Administrative Clerical Trainees, AAF | Grades | 2947 | 123.7 | 11.1 | .40 |
| Clerical Trainees, AAF | Grades (weighted) | 123 | 125.9 | 9.9 | .44 |
| Clerical Trainees, Armored | Grades | 119 | 125.3 | 8.3 | .33 |
| Clerical Trainees, MAAC | Grades | 199 | 116.8 | 12.0 | .62 |
| Airplane Mechanic Trainees | Grades | 99 | 104.8 | 10.6 | .32 |
| Airplane Mechanic Trainees | Grades | 3081 | 118.1 | 10.7 | .35 |
| Motor Mechanic Trainees | Grades | 318 | 88.3 | 24.4 | .69 |
| Tank Mechanic Trainees | Grades | 237 | 116.6 | 11.3 | .33 |
| Aircraft Armor Trainees | Grades | 1907 | 117.3 | 10.9 | .40 |
| Aircraft Armor Trainees | Ratings | 449 | 112.7 | 12.1 | .27 |
| Aircraft Welding Trainees | Grades | 583 | 114.8 | 10.3 | .26 |
| Bombsight Maintenance Trainees | Grades | 195 | 129.1 | 10.5 | .31 |
| Sheet Metal Trainees, AAF | Grades | 764 | 115.6 | 10.3 | .27 |
| Teletype Maintenance Trainees, AAF | Grades | 487 | 123.5 | 12.1 | .20 |
| Radio Operator & Mechanic Trainees, AAF | Grades | 1055 | 122.4 | 11.1 | .32 |
| Radio Operator & Mechanic Trainees, AAF | Code Reg Speed, WPM | 217 | 117.4 | 11.7 | .24 |
| Radio Operator Trainees, WAAC | Grades | 152 | 116.2 | 11.7 | .38 |
| Radio Mechanic Trainees, AAF | Grades | 419 | 108.0 | 13.0 | .49 |
| Gunnery Trainees, Armored | Grades | 66 | 120.0 | 12.1 | .50 |
| Field Artillery Trainees, Instrument and Survey | Grades | 68 | 102.7 | 6.5 | .33 |

(Continued)

TABLE 3. VARIOUS EXAMPLES OF VALIDITY COEFFICIENTS FOR AGCT (Continued)

| Population | Criterion | N | Mean | SD | r |
|---|---|---|---|---|---|
| Motor Transport Trainees, WAAC | Grades | 269 | 111.4 | 13.6 | .31 |
| Tank Driver Trainees | Ratings | 330 | 87.7 | 19.5 | .16 |
| Truck Driver Trainees | Road Test Ratings | 421 | 95.5 | 20.1 | .13 |
| Bombardier Trainees, AAF | Grades, Academic | 40 | 111.5 | 18.6 | .62 |
| Aircraft Warning Trainees, Plotter-Teller | Grades, Theory | 119 | 107.1 | 15.6 | .73 |
| Aircraft Warning Trainees, Plotter-Teller | Grades, Performance | 119 | 107.1 | 15.6 | .26 |
| Intelligence Trainees, AAF | Grades, Academic | 104 | 118.9 | 10.6 | .51 |
| Photography Trainees, AAF | Grades | 431 | 123.0 | 11.9 | .24 |
| Cryptography Trainees, AAF | Grades, Phase 1 | 417 | 129.9 | 9.7 | .31 |
| Weather Observer Trainees, AAF | Grades | 1042 | 130.2 | 12.5 | .43 |
| Officer Candidates, Infantry | Grades, Academic | 103 | 123.0 | 10.8 | .30 |
| Officer Candidates, Ordnance | Grades, Academic | 190 | 128.2 | 9.6 | .41 |
| Officer Candidates, Signal Corps | Grades, Academic | 213 | 128.6 | 10.1 | .36 |
| Officer Candidates, Tank Destroyers | Grades, Academic | 52 | 125.8 | 10.7 | .44 |
| Officer Candidates, Transportation Corps | Grades, Academic | 314 | 126.4 | 9.8 | .38 |
| Officer Candidates, WAAC | Grades, Academic | 787 | 128.4 | 11.3 | .46 |
| Officer Candidates, Infantry | Leadership Ratings | 201 | 122.6 | 10.8 | .12 |
| Officer Candidates, Ordnance | Leadership Ratings | 190 | 128.2 | 9.6 | .09 |
| Officer Candidates, 13 Arms and Services | Success vs Failure | 5186 | 128.7 | 10.0 | .28[a] |

(Continued)

TABLE 3. VARIOUS EXAMPLES OF VALIDITY COEFFICIENTS FOR AGCT (Continued)

| Population | Criterion | N | Mean | SD | r |
|---|---|---|---|---|---|
| AST Trainees, basic engineering | Grades, Inorganic Chemistry | 222 | 126.6 | 7.8 | .21 |
| AST Trainees, basic engineering | Grades, Math. (Trig.) | 222 | 126.6 | 7.8 | .16 |
| AST Trainees, personnel psychology | Ranks in Statistics | 132 | 134.2 | 10.4 | .25 |
| AST Trainees, personnel psychology | Ranks in Tests & Measurements | 130 | 134.0 | 10.3 | .29 |
| West Point Cadets, 4th Class | Grades, English[b] | 932 | 131.3 | 10.9 | .40 |
| West Point Cadets, 4th Class | Grades, Mathematics[b] | 932 | 131.3 | 10.9 | .43 |
| West Point Cadets, 4th Class | Grades, Military Topography | 932 | 131.3 | 10.9 | .40 |
| West Point Cadets, 4th Class | Grades, Tactics | 932 | 131.3 | 10.9 | .29 |
| West Point Cadets, 4th Class | Grades, French[b] | 167 | 130.2 | 11.0 | .22 |
| West Point Cadets, 4th Class | Grades, German[b] | 164 | 132.4 | 10.9 | .20 |
| West Point Cadets, 4th Class | Grades, Spanish[b] | 932 | 131.3 | 10.9 | .19 |
| West Point Cadets, 4th Class | Grades, Portuguese[b] | 168 | 130.0 | 10.3 | .12 |

Source: PRS Staff (1945), p. 767.

[a] Biserial Correlation.

[b] First Term.

Note, for example, in Table 3 that the four AGCT validities given for clerical courses ranged from r=.33 to r=.62 and that the six validities for the Officer Candidate courses ranged from r = .30 to r = .46 using grades as the criterion. Such results served to reinforce the perception of need for empirical validation of tests for each new application.

Of greatest significance was the aggregated success of the World War II research experience in providing applications for the work-place--a development considered by many as signaling the coming of age of psychology.

In 1949 the tests of the Army Classification Battery (ACB) were organized into aptitude areas, or combinations of tests for assigning individuals to various Military Occupational Specialties (MOS). The resulting classification system was a major innovation in military personnel operations. When compared with the single measure for the AGCT of World War II, tests developed with differential classification in mind were believed to meet more total personnel requirements with better overall validity. In the old system, using a single measure of general mental ability, individuals with high scores would be assigned to jobs demanding complex cognitive skills, while individuals with low scores would be assigned to less complex jobs. In the new system using aptitude area scores, classification would be based on demonstration of specific cognitive ability composites necessary for a particular job while at the same time utilizing total human resources more efficiently. Thus aptitude areas allowed the use of scores that indicated differences in the levels of abilities and differences <u>among</u> abilities <u>within</u> each individual (inter- and intra-individual differences).

The value of using several aptitude areas, rather than one composite, depends upon the presence of potential allocation efficiency (PAE) in the battery from which the tests comprising the aptitude areas were drawn. There was considerable PAE in the various versions of ACB during the first fifteen years of its use. Unpublished simulation studies conducted by the Army Research Institute showed a steadily declining trend in the amount of PAE present with each change of ACB content during the period that the ACB was being transitioned into the Armed Services Vocational Aptitude Battery.

## B. ARMED SERVICES VOCATIONAL APTITUDE BATTERY

In 1976, the Armed Services Vocational Aptitude Battery (ASVAB) was introduced for use by all military services as the common or joint-service selection and classification battery. The ASVAB essentially consisted of parallel forms of the subtests that comprised the Army Classification Battery of that period. In 1980, the then current versions of ASVAB (Forms 8/9/10) were introduced into operational use. This battery dropped or combined some of the old subtests and added a few subtests to form a new battery of ten subtests. (In 1984 parallel versions--Forms 11/12/13--were put into operational use.) The subtests of ASVAB (Forms 8/9/10) are shown in Table 4.

TABLE 4. ASVAB SUBTEST, TESTING TIMES AND RELIABILITIES

| Subtest | | Testing time (min) | Reliability |
|---|---|---|---|
| GS | General Science | 11 | .86 |
| AR | Arithmetic Reasoning | 36 | .91 |
| PC | Paragraph Comprehension | 13 | .81 |
| WK | Word Knowledge | 11 | .92 |
| NO | Numerical Operations | 3 | .78 |
| CS | Coding Speed | 7 | .85 |
| AS | Auto Shop Information | 11 | .87 |
| MK | Mathematical Knowledge | 24 | .87 |
| MC | Mechanical Comprehension | 19 | .85 |
| EI | Electronics Information | 9 | .82 |

Source: McLaughlin, Rossmeissl, Wise, and Brant (1984), p. 9.

Table 4 also gives testing times and reliabilities. Reliability estimates were based on a sample of 19,359 applicants for military service. Estimates for the eight power tests are KR-20 reliabilities and for the two speeded tests (NC and CO) alternate form reliabilities. Two subtests, PC and WK, are usually combined to form a general verbal ability subtest, VE. The ten subtests of ASVAB were combined into nine Army aptitude area composites shown in Table 5. The composites are used to assign individuals to various types of training programs or courses of instruction. Upon completion of training, individuals are assigned to specific Army jobs or MOS from among the 260 or so entry level Army MOS.

Individual MOS are clustered or grouped into a set of MOS families or career fields that are comparable to civilian job family taxonomies. Although the tests in the ASVAB and jobs exist within the military context, they were shown to be representative of the civilian setting (Hunter, Crosson, and Friedman, 1985). The nine job families encompass the spectrum of civilian jobs in the Dictionary of Occupational Titles. As shown in Table 5, a different aptitude area composite is used in assigning individuals to an MOS in each of the nine job families. Over the years, MOS clusters such as clerical/administrative or electronics repair were aggregated into such job families on the basis of judgment and empirical data. Jobs in a cluster were judged to have similar content and career ladders and also were demonstrated to require similar combinations of measured abilities. As new MOS were developed over the years, they were assigned to one of the existing job clusters on the basis of judgment and available data.

The Army Research Institute, in its Project A Study, is currently engaged in a remarkably ambitious, large-scale longitudinal effort designed to address many of the key scientific issues in selection and classification. A central aim of the research is to develop new types of predictors, and to validate

TABLE 5. OPERATIONAL COMPOSITES OF THE ASVAB IN USE BY THE ARMY IN 1984

| Cluster or<br>Job Family | Aptitude<br>Area<br>Composite | Subtests<br>Comprising<br>Aptitude Areas |
|---|---|---|
| Clerical/Administrative | CL | (VE+NO+CS)[a] |
| Combat | CO | (AR+CS+AS+MC) |
| Electronics Repair | EL | (GS+AR+MK+EI) |
| Field Artillery | FA | (AR+CS+MK+MC) |
| General Maintenance | GM | (GS+AS+MK+EI) |
| Mechanical Maintenance | MM | (NO+AS+MC+EI) |
| Operators/Food | OF | (VE+NO+AS+MC) |
| Surveillance/Communications | SC | (VE+NO+CS+AS) |
| Skilled Technical | ST | (GS+VE+MK+MC) |

[a] VE, general verbal ability, combines subtests PC and WK.

them, along with existing ASVAB tests, against specially developed job performance measures, rather than just against training grades or the usual ratings. (See Eaton, Hanser, and Shields, 1986, for a complete description of research goals.)

The validation results described below deal with ASVAB (Forms 8/9/10) aptitude area composites and are based upon an analysis of 71,000 individuals on whom training and job proficiency evaluations in the form of Skill Qualification Test (SQT) scores were available. Analyses based on training outcomes included 81 MOS; and 46 of the same MOS were included in both samples, i.e., MOS that had both training and SQT scores. The results reported are from a comprehensive ARI report, McLaughlin, Rossmeissl, Wise, and Brant (1984).

For purposes of analysis, criteria were partitioned into analysis "cells" of at least 100 cases each and standardized within each cell. For the training data, a criterion cell was defined as an MOS, school and course combination. For the SQT data, a criterion cell was defined as an MOS, Track (skill level), and SQT-year combination. Thus cases within cells were designed to be homogeneous with regard to all variables upon which each evaluation was based. Still these analyses were based on the largest single database yet obtainable for ASVAB validation. The training cells' median N was 224; the SQT cells median N was 263; and about one third of both training and SQT cells had an N of 500 or over.

For the training criteria, end-of-course test scores were obtained; for job proficiency criteria, SQT scores were used, based upon administration one year after the ASVAB was administered. SQTs have been used by the Army since 1977 to assess individual qualifications for promotions. At present there are SQTs for only about 100 of the 260 entry level MOS. Each year a separate SQT is contructed for each MOS and skill level within that MOS. The SQT measures a soldier's ability to perform tasks specified in the Soldier's Manual. A test may sample from 12 to 36 or more tasks and soldiers are allowed to prepare in advance for the tasks to be tested. A SQT may consist of both hands-on and multiple-choice job knowledge items. However, only results of multiple-choice job knowledge items were available for use in this study. No reliability estimates were available for the SQT or end-of-course scores.

Table 6 gives the validities for the operational aptitude area composites used for assigning individuals to MOS within job families. Data were weighted by the number of accessions in each MOS and the proportions of observations for an MOS in each criterion cell.

Three major points should be made concerning the results shown in Table 6. First, the effects of restriction in range are quite apparent. The corrected or adjusted mean validity is .10 correlational points higher than the sample validity for both criteria, the corrected validity better reflecting operational or true effectiveness.

Second, job proficiency is better predicted than training: the average validity of the aptitude areas for job proficiency was r = .47 and for training was r = .40. This reversal of the normal finding (or better prediction of training than job proficiency) might be partially attributable to the wholly paper-and-pencil format of the job proficiency criterion as measured by the SQT--as compared to the final course grade format which had some rating and hands-on components included along with paper-and-pencil class test measures. Of most significance, however, is that the operational end-of-course grades used in the validation study were based on mostly criterion-referenced scores, indexing successful mastery of needed skills, that were then being employed by the Army. Passing students were expected to perform nearly perfectly on tests. Thus grades lacked the discriminability and variance normally associated with norm-referenced final course grades that attempt to discriminate between high and low performers.

A more representative index of the norm-referenced type of Army aptitude area validities against final course grades is provided by Maier and Fuchs (1972). The mean validity of comparable aptitude areas of the Army Classification Battery, corrected for range restriction against nine comparable job families, was found to be r = .65 based on a sample of 25,000 individuals in over 100 different entry MOS.

Maier and Grafton (1981) also validated ASVAB 8/9/10 against course grades, SQTs, and combination of both types of criteria across the same nine job families used in the McLaughlin et al. (1984) study. They found validities for the job families of:

CL = .53 (mixture of SQT and training grades)
CO = .56 (SQT only)
EL = .59 (SQT only)
FA = .63 (SQT only)
GM = .76 (SQT only)
MM = .52 (mixture of SQT and training grades)
OF = .61 (SQT only) SC = .55 (training grades only)
ST = .55 (mixture of SQT and training grades)

TABLE 6. VALIDITIES OF OPERATIONAL APTITUDE AREA COMPOSITES FOR THE NINE ARMY JOB FAMILIES

| | | Training Performance Validity | | | Job Proficiency Validity | | |
|---|---|---|---|---|---|---|---|
| | Aptitude Area | N[a] | Sample | Corrected[b] | N[a] | Sample | Corrected[b] |
| Clerical/Administrative | CL | 5300 | .19 | .40 | 8000 | .29 | .49 |
| Combat | CO | 2900 | .25 | .36 | 16000 | .33 | .44 |
| Electronic Repair | EL | 2600 | .22 | .40 | 6000 | .28 | .45 |
| Field Artillery | FA | 1800 | .25 | .35 | 7000 | .34 | .45 |
| General Maintenance | GM | 1900 | .29 | .52 | 1300 | .23 | .40 |
| Mechanical Maintenance | MM | 5400 | .28 | .35 | 4300 | .28 | .45 |
| Operators/Food | OF | 4600 | .20 | .35 | 7700 | .33 | .50 |
| Surveillance/Communications | SC | 1500 | .18 | .34 | 3600 | .29 | .47 |
| Skilled Technician | ST | 3200 | .32 | .54 | 6900 | .32 | .55 |
| MEAN | | | .31 | .40 | | .37 | .47 |

Source: Adapted from McLaughlin et al. (1984).

[a] Rounded to the nearest hundred.

[b] For restriction in range.

The mean validity of ASVAB across all jobs was r = .60. Mean validities of r = .60 to r = .65 for training grades are similar to the range of validities typically found in validation studies in the other services.

Third, and of greatest significance, is the substantial high level of validity coefficients ior each operational aptitude area selector for its job family. It should be noted that the value of r = .47 against job proficiency, here measured as job knowledge, is considerably higher than the overall observed (uncorrected) validity range between r = .22 and r = .28 reported in previous major reviews of heterogenous collections of validity (Ghiselli, 1973; Boehm, 1982; and Schmitt, Gooding, Noe, and Kirsh, 1984.) The mean validity of r = .47 is of special significance because of the relatively substantial sample sizes and the homogeneity of the comparisons made. Consequently, this value, corrected only for restriction in range, may be one of the better characterizations in the literature of the operational effectiveness of a major aptitude battery against an objective measure of job proficiency.

Tables 7 and 8 show the corrected validities for restriction in range for each aptitude area composite across job families. In making corrections, it was assumed that explicit selection was made on all ASVAB subtests. This assumption would lower differential validity estimates if actual selection assignment was more explicit for some subtests than others. The validities were obtained by averaging the validities for the individual MOS within each aptitude area family and weighing by accession in each MOS. The main diagonals give the operational validities of the aptitude area associated with each job family.

For the training criterion, all validities range between r = .27 and r = .54. The operational composite tends to be the best selector or close to the best selector within a few correlational points, except for the CL aptitude area composite.

TABLE 7. AVERAGE CORRECTED TRAINING VALIDITIES OF APTITUDE AREA COMPOSITES FOR THE NINE ARMY JOB FAMILIES

| Job Family | | | | | | Aptitude Area Composite[a] | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | N[b] | CL | CO | EL | FA | GM | MM | OF | SC | ST | Average |
| Clerical/Administrative | 5300 | 40 | 43 | 45 | 46 | 42 | 39 | 42 | 42 | 45 | 43 |
| Combat | 2900 | 30 | 36 | 33 | 35 | 33 | 34 | 35 | 34 | 34 | 34 |
| Electronic Repair | 2600 | 35 | 42 | 40 | 41 | 39 | 40 | 41 | 39 | 40 | 40 |
| Field Artillery | 1800 | 27 | 27 | 34 | 35 | 35 | 37 | 36 | 32 | 33 | 34 |
| General Maintenance | 1900 | 42 | 52 | 51 | 50 | 52 | 52 | 52 | 49 | 50 | 50 |
| Mechanical Maintenance | 5400 | 33 | 44 | 42 | 41 | 44 | 44 | 44 | 40 | 42 | 42 |
| Operators/Food | 4600 | 28 | 35 | 34 | 33 | 35 | 34 | 35 | 33 | 35 | 34 |
| Surveillance/Communications | 1500 | 33 | 35 | 35 | 36 | 33 | 32 | 34 | 34 | 35 | 34 |
| Skilled Technician | 3200 | 46 | 52 | 53 | 51 | 52 | 50 | 53 | 51 | 54 | 51 |
| Average | | 35 | 41 | 40 | 43 | 40 | 39 | 40 | 38 | 40 | 40 |

Source: McLaughlin et al. (1984) p. 25.

[a] Decimals omitted.

[b] Rounded to the nearest hundred.

TABLE 8. AVERAGE CORRECTED JOB PERFORMANCE VALIDITIES OF APTITUDE AREA COMPOSITES FOR THE NINE ARMY JOB FAMILIES

| Job Family | | Aptitude Area Composite[a] | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | N[b] | CL | CO | EL | FA | GM | MM | OF | SC | ST | Average |
| Clerical/Administrative | 8000 | 49 | 52 | 55 | 55 | 51 | 48 | 52 | 51 | 55 | 52 |
| Combat | 16000 | 36 | 44 | 44 | 43 | 43 | 43 | 44 | 40 | 44 | 42 |
| Electronic Repair | 6000 | 35 | 45 | 45 | 43 | 45 | 44 | 45 | 41 | 45 | 43 |
| Field Artillery | 7000 | 36 | 46 | 46 | 45 | 46 | 46 | 46 | 42 | 46 | 44 |
| General Maintenance | 1300 | 33 | 41 | 40 | 40 | 40 | 40 | 41 | 38 | 41 | 39 |
| Mechanical Maintenance | 4300 | 32 | 44 | 43 | 41 | 45 | 45 | 44 | 39 | 43 | 42 |
| Operators/Food | 4700 | 40 | 51 | 51 | 48 | 51 | 49 | 50 | 46 | 51 | 49 |
| Surveillance/Communications | 3600 | 40 | 52 | 51 | 49 | 52 | 51 | 51 | 47 | 52 | 49 |
| Skilled Technician | 6900 | 48 | 54 | 55 | 55 | 53 | 51 | 54 | 52 | 55 | 53 |
| Average | | 39 | 48 | 48 | 47 | 47 | 46 | 47 | 44 | 48 | 46 |

Source: McLaughlin et al. (1984) p. 27.

[a] Decimals omitted.

[b] Rounded to the nearest hundred.

For the job proficiency criterion, all validities range between r = .32 and r = .55. Again the operational composite tends to be the best or close to the best selector within a few correlational points, except for the CL aptitude area composite. One MOS job family, Skilled Technician, is more predictable than the other job families for both criteria, and one, Combat, is less predictable than the others. Overall, then, all the aptitude areas are highly effective predictors and fall within a remarkably narrow range. (Later analyses identified improved predictor composites that were made operational in 1985 for the CL and SC job families.)

A serious shortcoming of the existing battery is its inability to differentiate among job families. The same aptitude area used to select individuals specific to an MOS within a job family does nearly as well for MOS in other job families. Each aptitude area is about as valid for other job families as it is for its own. While the operational composites are highly valid, the battery lacks differential validity.

It is generally assumed that the utility of the classificction process is a direct function of differential validity. More precisely, however, differential validity is the level of prediction of differences among criterion scores. It requires a simulation study to translate the effect of differential validity into utility. PAE is a direct expression of utility in terms of average predicted performance.

During the last two decades both test development and the selection of tests for inclusion in operational batteries have been directed toward the objective of maximizing the average validity of aptitude composites while ignoring the possibility that PAE might be lowered in the process. However, it might be possible to find and exploit the presence of PAE in future operational batteries designed expressly for that purpose and also retain the conventional ability domains present in the ASVAB (see section on PAE later in this report).

As noted earlier, the reliabilities of the two criteria were not known. However, on 11,000 of the same individuals in 81 MOS both end-of-course and SQT measures were available. This relationship is not commonly reported in the literature because of the practical difficulty of obtaining such data. High correlations between the two measures would have been an indication of high reliabilities; unfortunately that did not prove to be the case. Correlations, uncorrected for attenuation and range restriction, varied between r = .12 and r = .56 over the 81 MOS with a mean of r = .22. One interpretation of this finding is that end-of-course grades and SQT scores measured different facets of performance in this study. Yet training criteria are often considered surrogates for job performance criteria. The unreliabilities of course grades and SQT scores, and range restrictions might be equally likely explanation of the low correlations. Another explanation might be that the same basic constructs are being measured but that over time and experience individuals change their *rank order* in job skill and knowledge (Schmidt, Hunter, and Outerbridge, 1986).

Low correlations between the two criteria do not imply that the same ability tests will not predict both criteria. As shown in Tables 7 and 8, ability tests valid for one of the two criteria also were substantially valid for the other criterion.

A more comprehensive study of test criterion combinations would include other predictor constructs in addition to the ASVAB validated against carefully developed criterion measures of training, hands-on and job-knowledge performance measures, and performance ratings. The Army Research Institute has developed such predictor-criterion measures and is currently carrying out a validation study. Preliminary findings will be reviewed later in this report.

## C. AIRMAN SELECTION AND CLASSIFICATION BATTERIES

The U.S. Air Force has employed multiple aptitude batteries for selection and classification since the late 1940s. More than a dozen different operational batteries have been used from 1951 to 1974. Some batteries could be characterized as major changes, others as revisions. The underlying basis for the development of a classification battery in the Air Force is the same as for the Army described above--that success on each job can be associated with a specific pattern of abilities and that the most important abilities that are common across jobs can be identified and measured. Accordingly, an empirically determined composite of abilities in a classification battery can be used to predict performance in each job or job family. In practice all services use aptitude composites of subtests to predict success in clusters of jobs which have been judged or determined to be homogeneous.

Weeks, Mullins, and Vitola (1975) published an evaluation of the first ten operational classification batteries used by the Air Force since the end of World War II. Details relating to ASVAB-3, the latest battery in the Weeks et al. report to be evaluated, are abstracted below.

ASVAB-3 was first used operationally by the services in September 1973. The Air Force developed Aptitude Indices (AI) from the nine subtests of the ASVAB. The subtest comprising each AI are shown in Table 9. The 46 validity coefficients shown in Table 10 represent ASVAB-1 validities against technical school final course grades. The validities shown in Table 10, corrected for restriction in range, vary between r = .20 amd r = .87 with a median of r = .68. The correlations for similar AIs in ASVAB-1 and ASVAB-3 range from r = .72 to r = .83.

Kyllonen (1986) summarizes the validities for the ten batteries evaluated by Weeks et al. (1975) and adds validities from three later test batteries. Table 11 gives median validities for the thirteen test batteries in operational use for nearly 40 years. The validities shown are consistently high across all batteries and all courses. The results are quite congruent with results for the Army reported by Maier and Fuchs (1972).

TABLE 9. APTITUDE INDICES FOR ASVAB-3
DEVELOPED BY THE AIR FORCE, 1973

| | Aptitude Index[a] | | | |
|---|---|---|---|---|
| Subtest | M | A | G | E |
| Coding Speed | - | X | - | - |
| Word Knowledge | - | X | X | - |
| Arithmetic Reasoning | - | - | X | X |
| Tool Knowledge | X | - | - | - |
| Space Perception | - | - | - | X |
| Mechanical Comprehension | X | - | - | - |
| Shop Information | X | - | - | - |
| Automotive Information | X | - | - | - |
| Electronics Information | - | - | - | X |

Source: Adapted from Weeks, Mullins, and Vitola (1975), p. 43.

[a] M = Mechanical; A = Administrative; G = General; E = Electronics.

TABLE 10. ASVAB-1 APTITUDE INDICES VALIDITIES
(CORRECTED FOR RESTRICTION IN RANGE)

| Aptitude Index | Technical School Course | N | r |
|---|---|---|---|
| Mechanical | Aircraft Pneudraulic Repairman | 115 | .62 |
| | Aircraft Fuel Systems Mechanic | 66 | .29[a] |
| | Aircraft Maintenance Specialist (Reciprocating Engine) | 238 | .67 |
| | Aircraft Maintenance Specialist (Jet 1 and 2 engines) | 691 | .55 |
| | Aircraft Maintenance Specialist (Jet over 2 engines) | 302 | .63 |
| | Aircraft Maintenance Specialist (Turbo-prop) | 271 | .66 |
| | Jet Engine Mechanic | 485 | .61 |
| | Missile Mechanic | 53 | .67 |
| | Munitions Maintenance Specialist | 73 | .55 |
| | Weapons Mechanic | 345 | .53 |
| | Vehicle Repairman | 52 | .82 |
| | Air Frame Repair Specialist | 150 | .70 |
| | Corrosion Control Specialist | 51 | .71 |
| | Electrical Power Production Specialist | 120 | .64 |
| | Air Cargo Specialist | 50 | .55 |
| | Aircraft Loadmaster | 50 | .59 |

[a] Significant at the .05 level, all of the other validity coefficients are significant at p=.01 level.

(Continued)

TABLE 10. ASVAB-1 APTITUDE INDICES VALIDITIES (CORRECTED FOR RESTRICTION IN RANGE) (Continued)

| Aptitude Index | Technical School Course | N | r |
|---|---|---|---|
| Administrative | Communication Center Specialist | 215 | .64 |
| | Printer Systems Operator | 91 | .50 |
| | Morse Systems Operator | 84 | .57 |
| | Ground Radio Operator | 215 | .38 |
| | Inventory Management Specialist | 789 | .75 |
| | Disbursement Accounting Specialist | 122 | .37 |
| | Personnel Specialist | 262 | .86 |
| General | Imagery Interpreter Specialist | 116 | .86 |
| | Weather Observer | 99 | .84 |
| | Air Traffic Control Operator | 156 | .68 |
| | Aircraft Control and Warning Operator | 133 | .83 |
| | Medical Service, Fundamentals | 401 | .84 |
| | Medical Service Specialist | 50 | .84 |
| | Protective Equipment Specialist | 60 | .69 |
| | Fuel Specialist | 150 | .54 |
| | Security Specialist | 707 | .72 |
| Electronics | Aircraft Radio Repairman | 114 | .86 |
| | Aircraft Electronic Navigation Equipment Repairman | 138 | .82 |
| | Electronic Warfare Repairman | 62 | .82 |

(Continued)

TABLE 10. ASVAB-1 APTITUDE INDICES VALIDITIES
(CORRECTED FOR RESTRICTION IN RANGE) (Continued)

| Aptitude Index | Technical School Course | N | r |
|---|---|---|---|
| (Electronics) | Aircraft Inertial and Radar Navigation Systems Repairman | 71 | .85 |
| | Radio Relay Equipment Repairman | 61 | .85 |
| | Ground Radio Communications Equipment Repairman | 70 | .87 |
| | Electronic Communications and Cryptographic Equipment Systems Repairman | 50 | .64 |
| | Telecommunications Control Specialist/Attendant | 82 | .84 |
| | Weapons Control Systems Mechanic | 60 | .75 |
| | Communications and Relay Center Equipment Repairman (Electro/ Mechanical) | 52 | .69 |
| | Aerospace Photographic Systems Repairman | 66 | .59 |
| | Aerospace Ground Equipment Repairman | 208 | .83 |
| | Instrument Repairman | 68 | .67 |
| | Aircraft Electrical Repairman | 134 | .64 |

Source: Weeks, et al. (1975), p. 45.

TABLE 11. VALIDITY COEFFICIENTS OF AIR FORCE TEST BATTERIES

| Test Battery | Year | Number of Courses | Range of Validities | Median Validity | Sample Size |
|---|---|---|---|---|---|
| AC1-A | 1951 | 29 | .32 -- .77 | .61 | 261 |
| AC1-B | 1956 | 21 | .34 -- .77 | .60 | 402 |
| AC2-A | 1959 | 46 | .11 -- .80 | .57 | 124 |
| AQE-D | 1958 | 3 | .45 -- .50[a] | .47 | 182 |
| AQE-F | 1963 | 41 | .29 -- .90 | .63 | 433 |
| AQE-62 | 1962 | 4 | .75 -- .81[a] | .79 | 1493 |
| AQE-64 | 1968 | 57 | .38 -- .87 | .64 | 410 |
| AQE-66 | 1973 | 46 | .18 -- .90 | .68 | 115 |
| AQE-J | 1971 | 4 | .69 -- .84[a] | .82 | 3396 |
| ASVAB-3 | 1968 | 46 | .29 -- .87[a] | .68 | ---[c] |
| AQE/AFQT (1) | 1974 | 42 | .16 -- .63[b] | .42 | 1000 |
| AQE/AFQT (2) | 1974 | 43 | .16 -- .65[b] | .44 | 823 |
| AQE/AFQT (3) | 1974 | 57 | .22 -- .68[b] | .53 | 890 |

Source: Kyllonen (1986), p. 4.

NOTE. The first ten rows are adapted from Weeks, et al. (1975); the last three rows are adapted from Christal (1976).

[a] Inferred validities from test relationships with previous batteries for which actual validity studies were conducted.

[b] Not corrected for restriction of range.

[c] Unknown.

However, comparing the Air Force's results in Tables 10 and 11 with the Army's in Tables 6 through 8, it is clear that the overall Air Force's training validities were much higher: Air Force battery average validity of r = .65 for technical final course grades versus Army battery validity of r = .40 for final course grades and mean validity of r = .47 against job-knowledge tests. A possible explanation for the difference in the level of validities may relate to the technical content in the courses given by the two services. The Air Force courses tended to be more technical, specialized or cognitively complex than the Army courses, many of which, in addition, required much less formal classroom time. Ability tests generally are better able to predict more cognitively complex training courses and jobs than they do less cognitively complex training and jobs (Hunter, 1983b). Hunter, however, states that the validity of cognitive ability is expected to be high across all levels of complexity in training because cognitive ability predicts learning in all contexts. This assertion will be examined in more detail later. Shorter training time would probably result in lower reliability of trainin<sub></sub>j performance measures, and this in turn could be a reason for lower observed validities along with the lower complexity level.

A second, and more significant, explanation may be found in the distribution of end-of-course scores. As mentioned earlier, most of the end-of-course training scores used in the Army va]idation were criterion-referenced. Passing students were expected to perform nearly perfectly on relatively easy tests. Thus, the operational course grades used in the Army validation study described earlier were not designed to discriminate among students as would have been desired in a validation study.

Taken together, the Army and Air Force results clearly indicate the high level of effectiveness of tests in predicting training outcomes. However, as Weeks et al. (1975) point out, Air Force validity evaluations suffered from a lack of an empirical job performance criterion:

> The customary solution to this problem was to employ the available intermediate criterion, typically school course grades. As a result...the validity of the batteries for predicting successful job performance was an unknown, (p. 46).

We know from the research evidence cited for the Army above and the evidence cited in later sections of this report that the test-job correlations are substantial, even if the correlations between grades and job performance were found to be low. Additionally, technical grades are a measure of job knowledge at the end of training and job knowledge correlates with performance on job sample tests (Hunter, 1983a, Schmidt, Hunter, and Outerbridge, 1986, and Vineberg, 1982).

## D. ASVAB TRAINING VALIDITIES ACROSS THE SERVICES AND DIFFERENTIAL VALIDITY

ASVAB has been administered to high school students since 1968 for purposes of recruiting, vocational guidance and counseling. The ASVAB-14 currently being used for this program is a parallel form of ASVAB-8/9/10 and ASVAB-11/12/13. Four "factor composites" (verbal, quantitative, technical and speed) were being used until recently in reporting results to students and their schools. Currently three "academic" composites are used--verbal, math, and academic aptitude. Since ASVAB is believed to predict performance in civilian jobs as well as in military jobs, occupational composites also were developed to predict performance in four different job families. Hunter, Crosson, and Friedman (1985) evaluated the effectiveness of both the "factor composites" and the occupational composites on the extensive validity information now available on ASVAB-8/9/10. This evaluation provides an important additional source of data on the level of effectiveness of selection and classification tests across all military services.

Table 12 shows the average corrected validities for the occupational composites across all jobs in each service against training success criteria (final course grades). It should be noted, however, that for the Army comparison Hunter et al. (1985) used SQT scores obtained about a year after administration of the ASVAB. Validities obtained for the Army sample more properly should be considered as job proficiency validities against a job-knowledge criterion. The different nature of the criteria and differences in time between ASVAB and criterion testing reduces the interpretability of between-service comparisons. Mean validities were: Army, r = .48; Air Force, r = .74; Navy, r = .51; and Marines, r = .58. The overall mean across services, across all 190 jobs with a sample size of 103,700 was r = .58. The much higher level of Air Force validities against course grades, again, might be attributed to the more technical content or higher complexity of jobs in that service or to methodological differences in the criteria.

Hunter, Crosson, and Friedman (1985) drew a very significant conclusion after analyzing the occupational composite validities for each job family in each service. There were nine families in the Army, four in the Air Force, five in the Navy and six in the Marine Corps. The Marine Corps now has four composites and job families: CL, MM, EL, and GT. If different aptitudes predict different job families, the validity of each occupational composite should be highest for its own associated job family and lower for the other job families. Such a result would be indicative of differential validity. The results, unfortunately, indicated that each occupational composite is almost as valid for other job families as for its own. Similar results also can be seen in Table 12 by comparing validities across the four job families used in the high school program. The conclusion reached, then, is that the ASVAB composites provide high validity but little differential as predictors of training success across all jobs. With little or no PAE implied by the

## TABLE 12. AVERAGE TRAINING VALIDITIES OF ASVAB COMPOSITES[a] FOR FOUR JOB FAMILIES BY MILITARY SERVICE

| | | | Validity[b] | | | | |
|---|---|---|---|---|---|---|---|
| Service | Number of Jobs | Sample[c] | M&C | B&C | E&E | HS&T | Total |
| Army | 55 | 50,000 | .49 | .45 | .49 | .50 | .48 |
| Air Force | 70 | 29,700 | .70 | .74 | .77 | .74 | .74 |
| Navy | 31 | 7,600 | .50 | .49 | .53 | .53 | .51 |
| Marines | 34 | 16,400 | .58 | .58 | .53 | .61 | .58 |
| Total | 190 | 103,700 | .56 | .55 | .59 | .59 | .58 |

Source: Hunter, Crosson, and Friedman (1985). p. 116.

| [a]Job Family | ASVAB Subtests |
|---|---|
| M&C = Mechanical and Crafts | Arithmetic Reasoning + Mechanical Comprehension + Auto Shop Information + Electronics Information (AR+MC+AS+EI) |
| B&C = Business and Clerical | Word Knowledge + Paragraph Comprehension + Coding Speed + Mathematical Knowledge (WK+PC+CS+MK) |
| E&E = Electronics and Electrical | Arithmetic Reasoning + Mathematical Knowledge + Electronics Information + General Science (AR+MK+EI+GS) |
| HS&T = Health, Social and Technology | Word Knowledge + Paragraph Comprehension + Arithmetic Reasoning + Mechanical Comprehension (WK+PC+AR+MC) |

[b]Corrected for restriction in range.

[c]Rounded to the nearest hundred.

lack of differential validity, the benefits obtainable from using more than one occupational predictor composite would be negligible, if not zero.

## E. PREDICTION OF MILITARY JOB PERFORMANCE

Vineberg and Joyner (1982) summarized the literature published between 1952 and 1980 on predicting performance in military jobs. In a review of 114 studies, they found that 48 percent of the studies reported the use of ratings alone as a criterion and 30 percent used a measure of suitability (see below). In contrast, only 18 percent of the studies reported using an actual measure of job proficiency.

Vineberg and Joyner make the standard distinction between job proficiency and job performance, namely, contrasting what a person knows or can do with what a person actually does on the job. Proficiency usually is measured by a paper-and-pencil or a hands-on test of job tasks and is generally objective and reliable. Job performance measures, usually ratings, are generally subjective and less reliable than proficiency measures.

Correlations between written job-knowledge measures of proficiency and hands-on job sample measures of proficiency were generally found to be low, ranging from r = .00 to about r = .30. However, when job-knowledge tests were constructed, based only on information directly relevant to job performance, higher correlations were found, ranging from r = .58 to r = .78. The low reliability of ratings limited their relationship with other proficiency measure, with only a few correlations appearing above r = .30.

Hunter (1983a) also examined the relationships among the three types of criterion measures in a meta-analysis of 14 validation studies. His results were consistent with the Vineberg and Joyner results although the corrected correlations were higher, e.g., correlation between hands-on measures and ratings was r = .35 and correlation between hands-on measures and job knowledge was r = .67. The two sets of findings clearly indicate that each type of criterion measures different aspects of performance.

The suitability criterion, as employed in the Vineberg and Joyner review, is an index of overall adaption to military service. In most studies suitability is a composite criterion of two or more indices reflecting completion of term of enlistment, recommendation for reenlistment, advancement in grade or skill level, and performance ratings.

Table 13 shows the average validity coefficients for a variety of predictors against four criterion types. Predictors included ability tests, biodata, interest, personality, training grades among several others. Aptitude composites as used in this analysis were combined validities for operational selection or classification tests such as the Armed Forces Qualification Test, Army aptitude area scores, and Air Force aptitude indices. Validities also included several cross-validated experimental aptitude measures. The coefficients included for analysis were a mixture of validities, some corrected for restriction in range and others that were not.

TABLE 13. AVERAGE VALIDITIES OF VARIOUS PREDICTORS FOR FOUR TYPES OF CRITERIA

| Criterion | Number of Correlations | Median Validity |
|---|---|---|
| Job Knowledge | 110 | .40 |
| Task Performance | 18 | .31[a] |
| Global Rating | 204 | .15 |
| Suitability | 19 | .24 |

Source: Adapted from Vineberg and Joyner (1982), p. 8.

[a] More recent findings report validities for ASVAB against hands-on tests ranging from r = .56 to r = .59 in Marine Corps studies (Maier and Hiatt, 1984).

The validities given in Table 13 are consistent with rankings in most other studies: job-knowledge tests of proficiency are predicted best (r = .40), global ratings of performance are predicted least well (r = .15). The higher validity for job-knowledge criteria may be attributable to higher reliabilities for the criteria, to the fact that cognitive dimensions are shared between aptitude tests and job-knowledge tests, and to the relatively large number of aptitude composite validities present in the sample. In contrast, the lower validity of predictors against global ratings may be partially attributable to the low reliability of ratings and to the scarcity of cognitive components in job ratings. Vineberg and Joyner recommend that the:

> use of supervisors' ratings as the sole measure of job performance should be restricted to jobs for which motivation, social skills, and response to situational requirements are the only attributes worth measuring, (p. VIII.)

The validities of predictors against task performance (r = .31) and suitability (r = .24) criteria (similar to job potential) are sufficiently high to make them of considerable practical value in validation research.

Table 14 shows the validities of various types of predictors against a global rating of job performance and a suitability criterion. The rank order of predictive effectiveness for global ratings is consistent with other findings, training performance (r = .23) being the highest. However, the level of the validity coefficients found is lower than levels generally reported for civilian jobs, possibly pointing again to the low reliabilities for ratings, at least as obtained in the context of these studies. For the suitability criterion there is a different ordering of effectiveness of test types and level of validity reached, e.g., education was highest, r = .36.

TABLE 14. AVERAGE VALIDITIES OF VARIOUS TYPES OF PREDICTORS FOR GLOBAL RATINGS AND SUITABILITY

| Predictor Type | GLOBAL RATING | | SUITABILITY | |
|---|---|---|---|---|
| | Number of Correlations | Median Validity | Number of Correlations | Median Validity |
| Aptitude | 101 | .12 | 11 | .24 |
| Biographical Inventory | 12 | .17 | 4 | .29 |
| Education | 25 | .12 | 10 | .36 |
| Interest | 15 | .12 | -- | -- |
| Training Performance | 59 | .23 | -- | -- |
| Age | -- | -- | 10 | .21 |
| Ratings | -- | -- | 7 | .29 |

Source: Adapted from Vineberg and Joyner (1982) p. 14.

Considering all of the results thus far for the Army ASVAB aptitude areas, the Air Force aptitude indices, ASVAB validities across services, and the job validation review, the following conclusions are reached concerning predictive effectiveness of military selection and classification batteries:

1. Aptitude area composites or ability tests are highly effective predictors of technical training. Aptitude composites predict success in training in the mean range of r = .55 to r = .74, with means tending to cluster around r = .60. The more technical or cognitively complex the training, the higher the validity tends to be.

2. Aptitude composites predict job proficiency as measured by job-knowledge tests with a mean validity of about r = .47; aptitude composites predict hands-on performance with means clustering around r = .55 (Maier and Hiatt, 1984).
3. Aptitude composites predict global ratings of performance at r = .21 validity level, suitability indices at r = .35 and all ratings at r = .35 validity level, after correction for criterion unreliability of Vineberg and Joyner's (1984) data by Hunter and Hunter (1984, p. 85).

The validities reported thus far for aptitude composites against training criteria and job knowledge measures or proficiency criteria were based on large samples and large numbers of independent correlations across the spectrum of military jobs and represent the best estimates extant of operational or true effectiveness of aptitude composites in the military, when conventionally obtained performance measures are used. The magnitude of ASVAB validities will be reevaluated later in this report when an array of specifically and carefully developed, high quality job performance measures are employed as criteria (see the section on ASVAB validation using multiple criteria).

## F. TEST VALIDITIES ACROSS CIVILIAN JOBS

The validity results summarized chus far were for specific aptitude tests or test composites within the military context. We must turn to the classic work of Edwin Ghiselli (1966, 1973) to obtain a comprehensive summary of general trends in test validity within the civilian context. Starting in the 1920s, Ghiselli analyzed an enormous amount of validity information spanning a 50-year period. His summary provided simple and concise summaries of average validities for 20 different test types for predicting training and job proficiency in 21 different job families.

Table 15 is an adaptation of Ghiselli's summary tables useful for comparison to the Hunter and Hunter summaries (see Hunter and Hunter, 1984, below). Table 15 provides validities for four test types (cognitive, perceptual, psychomotor, and personality) across nine job families. The job families have been rearranged by Hunter (1981) according to Fine's (1955) scaling of cognitive complexity of jobs (see GATB below).

Ghiselli reported that the grand average validity across all tests and all jobs is r = .39 for the training criterion and r = .22 for the job proficiency criterion. Nevertheless, considering the totality of Ghiselli's results, the practical usefulness of ability tests as predictors of job performance is clearly demonstrated. Personality and interest measures are less useful except perhaps for managerial and sales jobs. Ghiselli writes, however, that for every job at least one type of test exists which has at least moderate validity. If for each job the highest average validity is considered, then the overall average of these maximal validities is r = .45 against training criteria and r = .35 against job proficiency criteria.

Ghiselli points out that the level of validities found for aptitude tests is quite respectable:

> Considering the considerable differences in the times when the investigations summarized here were performed, together with the large differences in the nature of the organizations in which they were conducted, and marked variations among the samples in such factors as age, sex, education, and background, the average validity coefficients presented here can be said to have a good deal of generality. Furthermore, since most of them are based upon a number of separate and distinct determinations they have a substantial measure of dependability and meaningfulness, (p. 475).

Ghiselli concludes that the validity values given are conservative and that judiciously selected combinations of tests would increase validities.

TABLE 15. MEAN VALIDITIES OF VARIOUS PREDICTORS FOR NINE JOB FAMILIES (GHISELLI)

| Job Family | Training Validity | | | | Job Proficiency Validity | | | |
|---|---|---|---|---|---|---|---|---|
| | Cog | Per | Mot | Person | Cog | Per | Mot | Person |
| Managerial | .29 | .23 | .25 | .53 | .25 | .25 | .14 | .22 |
| Clerical | .41 | .40 | .14 | .17 | .23 | .29 | .16 | .22 |
| Salesperson | — | -- | — | -- | .27 | .23 | .16 | .30 |
| Protective | .39 | .30 | — | -.11 | .20 | .21 | .14 | .21 |
| Service | .37 | .25 | .21 | — | .20 | .10 | .15 | .16 |
| Trades & Crafts | .41 | .35 | .20 | .16 | .24 | .24 | .19 | .24 |
| Industrial | .38 | .20 | .28 | — | .21 | .20 | .22 | .26 |
| Vehicle Operators | .25 | .09 | .31 | — | .18 | .17 | .25 | .26 |
| Sales Clerk | — | — | — | -- | .06 | -.02 | .09 | .35 |

Source: Adapted from Ghiselli (1973), pp. 468-476.

Note. Cog = general cognitive ability;
Per = general perceptual ability;
Mot = general psychomotor ability;
Person = personality tests.

## G. REANALYSIS OF GHISELLI'S OCCUPATIONAL VALIDITIES

Hunter and Hunter (1984) provide a reanalysis of Ghiselli's summary of aptitude test validities according to job complexity or information processing requirements, corrected for criterion unreliability and range restriction, and also provide multiple correlations for combinations of ability test types.

Table 16 shows corrected aptitude test validities for job proficiency criteria. The validities range from r = .20 to r = .61 for the three ability categories (general cognitive ability, general perceptual ability, and general psychomotor ability). A striking pattern is apparent: cognitive ability validities decrease systematically with decreasing job complexity (with the exception of validities for sales clerk) while psychomotor ability validities systematically increase with decreasing job complexity. Psychomotor tests tend to have their highest validities for job families where cognitive tests tend to have their lowest validities. Consequently, multiple correlations are quite high for all job families, ranging from R = .43 to R = .62, except for sales clerks where the multiple correlation is R = .28. These findings indicate a strong moderating effect of job complexity on cognitive and psychomotor ability validities.

Ghiselli's overall average observed validity coefficient for all tests across all jobs was r = .22. This validity of r = .22 was increased in the reanalysis to an average multiple correlation of R = .48. As Ghiselli noted, the validity of r = .22 is clearly an underestimate of overall operational effectiveness of aptitude tests. The value of R = .48 is a much more accurate summary value of the operational predictive power of aptitude tests.

## H. GENERAL APTITUDE TEST BATTERY (GATB)

Over the last 40 years, the U.S. Employment Service has validated the same test battery (GATB) in 515 studies, typically

TABLE 16. HUNTER'S REANALYSIS OF GHISELLI'S OCCUPATIONAL JOB PROFICIENCY VALIDITIES

| | Mean Validity | | | Beta Weight | | |
|---|---|---|---|---|---|---|
| Job Families | Cog | Per | Mot | Cog | Mot | Multiple R |
| Manager | .53 | .43 | .26 | .50 | .08 | .53 |
| Clerk | .54 | .46 | .29 | .50 | .12 | .55 |
| Salesperson | .61 | .40 | .29 | .58 | .09 | .62 |
| Protective professions worker | .42 | .37 | .26 | .37 | .13 | .43 |
| Service worker | .48 | .20 | .27 | .44 | .12 | .49 |
| Trades and crafts worker | .46 | .43 | .34 | .39 | .20 | .50 |
| Elementary industrial worker | .37 | .37 | .40 | .26 | .31 | .47 |
| Vehicle operator | .28 | .31 | .44 | .14 | .39 | .46 |
| Sales clerk | .27 | .22 | .17 | .24 | .09 | .28 |

Source: Hunter and Hunter (1984), based on Hunter (1981).

Note. Cog = general cognitive ability;
Per = general perceptual ability;
Mot = general psychomotor ability;
R = multiple correlation.
Mean validities have been corrected for criterion unreliability and for range restriction using mean figures for each predictor from Hunter (1980a) and King, Hunter, and Schmidt (1980).

using job knowledge tests as measures of training success and ratings as measures of job proficiency. The use of a set of uniform predictors and criteria is more similar to the military studies than it is to the quite varied studies assembled by Ghiselli.

The GATB consists of 12 tests that are combined into nine aptitude composites, shown in Table 17. The aptitudes in turn are grouped into three general ability factors: cognitive ability (general, verbal and numerical); perceptual ability (spatial visualization, pattern recognition and form perception); and psychomotor abilities (motor coordination, finger dexterity and manual dexterity).

TABLE 17. THE U.S. EMPLOYMENT SERVICE GENERAL APTITUDE TEST BATTERY AND APTITUDE COMPOSITES

| Symbol | Aptitude Composites | Test |
|---|---|---|
| G | General Intelligence | Vocabulary + Arithmetic Reasoning + Three Dimensional Space |
| V | Verbal Aptitude | Vocabulary |
| N | Numerical Aptitude | Computation + Arithmetic Reasoning |
| S | Spatial Aptitude | Three Dimensional Space |
| P | Form Perception | Tool Matching + Form Matching |
| Q | Clerical Perception | Name Comparison |
| K | Motor Coordination | Mark Making |
| F | Finger Dexterity | Assemble + Disassemble |
| M | Manual Dexterity | Place + Turn |

Source: Hunter (1983b) p. 17.

Table 18 shows the mean validities of GATB for training and job proficiency criteria at five levels of job complexity reported by Hunter (1983b). Jobs are clustered into job families on the basis of complexity rather than on task similarity. Hunter's method of ordering is based on Fine's (1955) Functional Job Analysis dimension scheme for rating people, data and things. As mentioned earlier, Fine's approach also was used in the reanalysis of Ghiselli's data given in Table 16. Of the 515 validation studies 90 used criteria of training success and 425 used criteria of job proficiency. The average sample size of the studies was 75. The 515 jobs were considered representative of the entire work force job spectrum. Validities were corrected for range restriction and attenuation, with a reliability of r = .80 assumed for the training criterion of job knowledge and r = .60 assumed for the job proficiency criterion of ratings.

Table 18 shows the average validity for cognitive ability to be r = .55 for training success and r = .45 for job proficiency--consistent with the general finding of higher validities of cognitive abilities for training criteria than for job proficiency criteria. The average validity for psychomotor ability is r = .26 for training success as compared to the much higher validity of r = .37 for job proficiency.

From Table 18 it can be seen that the validity of cognitive ability for job proficiency decreases from .56 to .23 with decreases in job complexity, and conversely psychomotor ability validity increases with decreasing job complexity. Results of cognitive ability for training, however, show a fairly high validity regardless of job complexity, although the validity of r = .65 for the highest complexity job level was much higher than the validity for the lower job complexity levels, with validities ranging from r = .50 to r = .57. Psychomotor ability results for training again show an increase of validity as a function of decreasing job complexity. Hunter states that the

TABLE 18. MEAN VALIDITIES[a] OF GATB FOR TRAINING AND JOB PROFICIENCY AT FIVE LEVELS OF JOB COMPLEXITY

| | | Mean validity for training success | | | | Mean validity for job proficiency | | | |
|---|---|---|---|---|---|---|---|---|---|
| Job Complexity | % U. S. workers | GVN | SQP | KFM | Multiple R | GVN | SPQ | KFM | Multiple R |
| Setting up | 2.5 | .65 | .53 | .09 | .65 | .56 | .52 | .30 | .59 |
| Synthesizing/coordinating | 14.7 | .50 | .26 | .13 | .50 | .58 | .35 | .21 | .58 |
| Analyzing/compiling/computing | 62.7 | .57 | .44 | .31 | .59 | .51 | .40 | .32 | .53 |
| Comparing/copying | 17.7 | .54 | .53 | .40 | .59 | .40 | .35 | .43 | .50 |
| Feeding/offbearing | 2.4 | — | — | -- | -- | .23 | .24 | .48 | .49 |
| Mean | | .55 | .41 | .26 | .57 | .45 | .37 | .37 | .53 |

Source: Adapted from Hunter (1983b), pp. 32-39.

Note. Dimensions and complexity families from Fine (1955).
GVN = cognitive ability;
SPQ = perceptual ability;
KFM = psychomotor ability;

GATB = General Aptitude Test Battery.

[a] Corrected for range restriction and attenuation.

findings for training are consistent with the need for good psychomotor abilities for hands-on training situations. On the other hand, cognitive ability increases in validity as a predictor of job proficiency as job demands become increasingly complex. Similarly, jobs that have low cognitive demands have a significant psychomotor demand. (Of course, there are jobs that are exceptions to this inverse relationship.) Taken as a whole, job complexity shows a strong effect on validity.

The complementary patterns of cognitive and psychomotor abilities lend themselves to various ability combinations or multiple correlations. Table 18 shows that the average multiple correlation for training is R = .57 as compared to an average of r=.55 for cognitive ability alone. Regression equations for computing the multiple correlations generally included only combinations of cognitive and psychomotor ability, and generally excluded perceptual ability. The average multiple correlation for training R = .57 and for job proficiency of R = .53 are impressively high levels of validities.

These values represent a good estimate of the predictive value of ability tests in the workplace. The validity of other predictor types needs to be contrasted with the validity of ability tests both to evaluate relative value and to seek possible increases in overall validity through combining tests. Such comparisons are made below.

## I. ALTERNATIVE SELECTION PROCEDURES

Reilly and Chao (1982) surveyed published and unpublished research during the 1970s concerning the validity of alternatives to conventional tests for employee selection. A conventional test was defined as a standardized measure of aptitude, knowledge, ability, personality or performance, with explicit administrative and scoring procedures. Work samples and assessment centers were not included as alternative employee selection procedures.

Alternative selection procedures were grouped into eight types of predictors or categories shown in Table 19. Average validities within each category were computed by weighting the Fisher Z transform of each validity coefficient by its sample size and then dividing by the total sample. Criteria employed were generally supervisory job ratings, tenure, productivity and salary. For biodata and peer evaluations, training grades in the military were also used.

TABLE 19. MEAN VALIDITIES FOR VARIOUS ALTERNATIVE PREDICTORS

| Predictor Type | Number of Correlations | N | Average Validity[a] |
|---|---|---|---|
| Biographical information | 44 | 11,600 | .35 |
| Interview | 12 | 1,000 | .19 |
| Peer evaluation | 33 | 12,800 | .41 |
| Self-assessment | 3 | 500 | .15 |
| Reference check | 10 | 5,700 | .14 |
| Academic performance | 20 | 2,700 | .20 |
| Expert judgment | 9 | 1,300 | .17 |
| Projective techniques | 5 | 300 | .18 |

Source: Extracted from Reilly and Chao (1982), pp. 1-61.

[a] Weighted Fisher Z transform.

Table 19 shows that only biographical inventories (r = .35) and peer evaluations (r = .41) have substantial validities. The level of these two alternatives, however, does not approach the average multiple correlation of ability combinations (e.g., R = .53) reported by Hunter and Hunter (1984). All other predictor types have a range of validities (r = .14 to r = .20) much below the levels generally obtained for ability tests. While the biodata variable is appealing because it is readily available in the employment setting, there are a number of constraints in its use including attenuation over time, i.e., biodata keys lose validity in follow-up studies, keys tend to be specific to the organization in which they were developed, keys require a very large sample in development to reduce chance selection of items, and finally accuracy and honesty of responses by examinees are open to question.

The use of peer ratings for job selection presents a number of practical and technical problems in the organizational setting. Peer ratings cannot be used for entry level jobs since applicants must be already trained and applicants' work needs to be known by a number of raters (as is more likely the case in the military setting). Additionally, peer evaluations are not readily accepted by many organizations and, when collected in the employment setting, require standardization of conditions. Reilly and Chao's data indicate strong support for the use of peer ratings for predicting subjective criteria and less support for predicting objective (verifiable) criteria. Peer evaluations seem to work best for supervisory and sales jobs.

Reilly and Chao conclude that of the alternatives reviewed, only biodata and peer evaluations have high validity. Additionally, situational interviews, miniaturized training tests, and unassembled examinations offer promise, but require considerable data.

## J. META-ANALYSES OF VALIDITIES

Schmitt, Gooding, Noe, and Kirsh (1984) reviewed all criterion-related validity studies published in Personnel Psychology and Journal of Applied Psychology from 1964 through 1982. In a number of meta-analyses they examined such effects on observed validity coefficients by sub-groups: of criteria used; job families; predictor types; and predictor-criterion combinations. A total of 99 articles produced 366 "summary" validity coefficients. In averaging the validity coefficients, each of the coefficients was weighted by its sample size. The variance of the coefficient was computed as well as the variance due to sampling error. The average observed validity coefficients reported were not corrected for range restriction or criterion unreliability.

Table 20 shows the average validity coefficients over six job families. The overall observed validity of r = .28 is roughly comparable to the classic result of r = .22 obtained by Ghiselli (1973). Large differences in the level of validities of different job families were found, sales and skilled labor coefficients being below r = .20 and the other coefficients above r = .30. Schmitt et al. state that sales and skilled labor jobs frequently involved the use of personality measures as predictors and turnover as a criterion--both generally associated with lower validity coefficients. Only eight percent of the overall variance in the validity coefficients was found to be accounted for by sampling error (a far smaller percentage than generally reported in validity generalization work). However, McDaniel, Schmidt, Raju, and Hunter (1986) pointed out that focusing on the percent variance accounted for could be misleading since (all else being equal) the larger the sample size, the lower the percent variance becomes due to sampling error. In fact, Schmitt et al. results were found by McDaniel et al. to be comparable with other validity generalizations, when corrections were made for unreliability and range restriction. Nevertheless, although the absolute amount of variance was about the same as in other such generalization studies, a higher proportion of variance remained.

Schmitt et al. point out that these sub-group averages involve a wide variety of test-criterion relationships and thus this finding was not unexpected.

TABLE 20. AVERAGE VALIDITY COEFFICIENTS AND STANDARD DEVIATIONS FOR VARIOUS JOB FAMILIES

| Job Family | Number of Correlations | Sample Total[a] | Average Validity | SD of Validity[b] |
|---|---|---|---|---|
| Professional | 81 | 18600 | .32 | .15 |
| Managerial | 93 | 43200 | .34 | .14 |
| Clerical | 36 | 9700 | .39 | .15 |
| Sales | 50 | 31700 | .17 | .09 |
| Skilled labor | 46 | 37700 | .18 | .12 |
| Unskilled labor | 60 | 9200 | .31 | .08 |
| TOTAL | 366 | 233400 | .28 | .13 |

Source: Adapted from Schmitt et al. (1984), p. 413.

[a] Rounded to nearest hundred.

[b] Observed values.

Table 21 gives the average validity coefficients for eight types of predictors. Validities ranged between r = .43 for supervisor/peer ratings (here used as predictors rather than criteria) to r = .15 for personality predictors. Note that the average validity found here for general mental ability, r = .25, is lower than for most other types of predictors. These findings which differ significantly from other validity generalization findings, will be discussed in conjunction with Hunter and Hunter's (1984) findings below.

Average validity coefficients for various criterion types are shown in Table 22. Validities ranged from r = .40 for work samples to r = .21 for productivity. Schmitt et al. point up the continuing concern with the use of ratings because of their low reliabilities and sensitivity to various biases. The performance rating is, of course, the most frequently used criterion measure, but the average performance rating validity of r = .26 is much lower than average validities for work samples, wages and status change.

TABLE 21. AVERAGE VALIDITY COEFFICIENTS AND STANDARD DEVIATIONS FOR VARIOUS TYPES OF PREDICTORS

| Predictor | Number of Correlations | Sample Total[a] | Average Validity | SD of Validity[b] |
|---|---|---|---|---|
| Special aptitude | 31 | 4300 | .27 | .14 |
| Personality | 62 | 23400 | .15 | .11 |
| Gen. mental ability | 53 | 40200 | .25 | .14 |
| Biodata | 99 | 58100 | .24 | .14 |
| Work sample | 18 | 3500 | .38 | .11 |
| Assessment center | 21 | 15300 | .41 | .05 |
| Super./peer evaluation | 31 | 6600 | .43 | .17 |
| Physical ability | 22 | 3100 | .32 | .22 |
| TOTAL | 366 | 233400 | .28 | .13 |

Source: Adapted from Schmitt et al. (1984), p. 415.

[a] Rounded to nearest hundred.

[b] Observed values.

TABLE 22. AVERAGE VALIDITY COEFFICIENTS AND STANDARD DEVIATIONS FOR VARIOUS TYPES OF CRITERIA

| Criterion | Number of Correlations | Sample Total[a] | Average Validity | SD of Validity[b] |
|---|---|---|---|---|
| Performance ratings | 140 | 17600 | .26 | .17 |
| Turnover | 48 | 12700 | .25 | .11 |
| Achievement/grades | 43 | 7200 | .27 | .20 |
| Productivity | 30 | 14900 | .21 | .08 |
| Status change | 46 | 52700 | .36 | .11 |
| Wages | 33 | 5500 | .38 | .15 |
| Work samples | 24 | 8200 | .40 | .16 |
| TOTAL | 366 | 233400 | .28 | .13 |

Source: Adapted from Schmitt et al. (1984), p. 416.

[a] Rounded to nearest hundred.

[b] Observed values.

Schmitt et al. performed additional analyses for various predictor-criterion combinations. Performance ratings are best predicted by assessment centers (r = .43) and supervisory/peer evaluations (r = .32) and work samples (r = .32) also are good predictors. Paper-and-pencil tests have lower validities: general mental ability (r = .22); personality (r = .21); and special aptitude (r = .16).

Some major conclusions drawn by Schmitt et al. are:

- Substantial levels of validity are found for predicting work success in all job families.
- Performance ratings yield lower levels of validity than do more objective criteria.

- Sample size variability accounts for about 25 percent of validity coefficient variance in this collection as compared to 50 to 100 percent reported for predictor-criterion sub-groups in other studies. (See McDaniel et al. 1986, mentioned earlier.)
- Various predictor types such as work samples and assessment centers are found to be superior to ability as predictors.

It is important to note that Schmitt et al. average observed (uncorrected) validities tend to be higher than comparable values found in validity generalization studies reported by Schmidt, Hunter and colleagues. The Schmitt et al. meta-analysis could have been biased because it included only published studies appearing in Personnel Psychology and Journal of Applied Psychology from 1964 to 1982, a period during which both journals accepted only validity studies that were unusual or remarkable in some way. In contrast, over half the studies included in Schmidt and Hunter's analyses were unpublished. Thus Schmitt et al. may have included studies which are less representative, but methodologically superior.

Schmitt et al. state that their results are not consistent with Hunter and Hunter's (1984) conclusions that cognitive tests are superior to other types of predictors (see section below). They also suggest an explanation similar to the one given here for the disparities between the findings. They write that much of the published research included in their analyses focused on the development and evaluation of alternative predictors. Ability tests were included only as standards for comparison or were merely "available" rather than being carefully developed standardized measures.

While Schmitt et al. conclusions are valid for this heterogenous collection, for a more detailed comparison of tests in which the criterion of job performance was the same and in which the type of use made of predictors was also the same, we turn to the results of Hunter and Hunter below.

## K. META-ANALYTIC COMPARISONS OF ALTERNATIVE PREDICTORS

Hunter and Hunter (1984) compared various predictor types using supervisory ratings as the criteria. It is important to note that two sets of comparisons were made. One set compared predictors that could be used for entry-level jobs where training would occur after hiring. A second set compared predictors used for promotion or certification where current performance on the job had been the basis for selection. This distinction is not made in most other meta-analytic comparisons despite the difference in purpose between the two in selection procedures and the quite disparate results obtained concerning the relative value of predictor types and the magnitude of observed validities.

Table 23 shows mean validities of 11 predictors suitable for use across entry level jobs. The validities, arranged in descending order, have been corrected only for criterion unreliability, except the ability composite. The ability composite validity was obtained from Hunter's (1983b) research on the GATB and, as described earlier, is a multiple correlation (combining cognitive and psychomotor abilities) corrected for both unreliability and range restriction. The four best predictors are ability composites (R = .53), job tryout (r = .44), biodata (r = .37) and reference check (r = .26). The validities for the next six highest predictors (e.g., experience, interview, training/experience rating, academic achievement, education and and interest) varied between r = .10 and r = .18, considerably lower than for ability. The only predictor with a validity that approached zero was age.

The mean validities across jobs for six predictors suitable for use in promotion or certification situations is given in Table 24. These validities were corrected only for criterion unreliability, except for the ability composite. Examinees being considered for promotion or certification were essentially performing the same kind of work in their current positions.

## TABLE 23. MEAN VALIDITIES[a] AND STANDARD DEVIATIONS OF VARIOUS PREDICTORS FOR ENTRY-LEVEL JOBS

| Predictor | Number of Correlations | Sample Total[b] | Mean Validity | SD |
|---|---|---|---|---|
| Ability composite | 425 | 32100 | .53 | .15 |
| Job tryout | 20 | --- | .44 | -- |
| Biography inventory | 12 | 4400 | .37 | .10 |
| Reference check | 10 | 5400 | .26 | .09 |
| Experience | 425 | 32100 | .18 | -- |
| Interview | 10 | 2700 | .14 | .05 |
| Training/exper. ratings | 65 | --- | .13 | -- |
| Academic achievement | 11 | 1100 | .11 | .00 |
| Education | 425 | 32100 | .10 | -- |
| Interest | 3 | 1800 | .10 | .11 |
| Age | 425 | 32100 | -.01 | -- |

Source: Adapted from Hunter and Hunter (1984), p. 90.

[a] Corrected for criterion unreliability.

[b] Rounded to nearest hundred.

TABLE 24. MEAN VALIDITIES[a] AND STANDARD DEVIATIONS OF PREDICTORS TO BE USED FOR PROMOTION OR CERTIFICATIONS

| Predictor | Number of Correlations | Sample Total[b] | Mean Validity | SD |
|---|---|---|---|---|
| Work sample test | -- | --- | .54 | -- |
| Ability composite | 425 | 32100 | .53 | .15 |
| Peer ratings | 31 | 8200 | .49 | .15 |
| Behav. consis. exper. rat. | 5 | --- | .49 | .08 |
| Job knowledge test | 10 | 3100 | .48 | .08 |
| Assessment center | -- | --- | .43 | -- |

Source: Adapted from Hunter and Hunter (1984), p. 91.

[a] Corrected for criterion unreliability.

[b] Rounded to nearest hundred.

Consequently, all the predictors in this category, except ability, are either ratings of current job performance or proficiency measures associated with the current job. Hunter and Hunter point out that excepting ability, these measures predict future performance based on measures of present or past performance. The mean validities ranged from r = .54 to r = .43, the best predictor being the work sample test (r=.54) followed very closely by the ability composite (R = .53).

I wish to comment on the disparate findings of Hunter and Hunter and Schmitt et al. concerning the validity of ability tests. In the comparisons of predictors by Hunter and Hunter average ability composite validity is based on corrected (for criterion unreliability and range restriction) multiple correlations of cognitive and psychomotor abilities combinations. On the other hand, Schmitt et al. give average uncorrected validities of single tests. Additionally, the Hunter and Hunter ability validities are based on the same carefully developed and standardized GATB tests, whereas the Schmitt et al. validities are based on a heterogenous assortment of general mental ability-job combinations. The sub-grouping in the Hunter and Hunter study takes into account the suitability of use of the test (entry-level vs. promotion); the Schmitt et al. study does not. In making comparisons among predictor types, the Hunter sub-grouping is more meaningful for use in the employment setting. In summary, then, when the validity of various predictor types are compared against the ubiquitous supervisory ratings, the results of Hunter and Hunter (1984) appear to be the current authoritative source.

An additional question of significance raised by Hunter and Hunter (1984) concerns the value of combining other types of predictors with ability. They point out that mathematical formulations limit the increase in validity due to the addition of a second predictor to the square of its validity at most. For example, adding the interview with a validity of r = .14 to ability with a validity of r = .53 in a multiple regression equation increases the validity, at most, from R = .53 to R = .55. Not surprisingly, a second predictor used with ability is given a small weight when optimally weighted. In practice, many users weigh predictors equally, not in proportion to validity, and thus the actual validity of the selection composite is lower than the validity of the best single test. Hunter and Hunter state that currently there are too few studies that consider different predictor types together to determine the value of combinations directly from meta-analyses. Hunter and Hunter do not consider the situation in which criteria may be differentially multidimensional across jobs or job families as may be the case in military classification. However, one major validation study does provide findings on combining different predictor types against different job performance criteria. We turn to this significant study next.

## L. ASVAB VALIDATION USING MULTIPLE CRITERIA

As mentioned earlier, the Army Research Institute is currently conducting a ten-year large scale research program to improve the selection, classification and assignment systems for Army enlisted personnel. The research program is now beginning its fifth year and has just completed initial validation of existing and experimental measures against major job performance dimensions for soldiers in their first tours of duty. Performance measures will be collected again during a second tour of duty.

The research program is divided into two major parts. Project A is concerned with the development of new measures of job performance that can be used as criteria against which to validate existing and experimental selection/classification measures; development of a utility scale for different performance levels across MOS; and the estimation of the relative effectiveness of alternative predictors in terms of their validity and utility. Project B is concerned with the development of computer-based decision aids to optimize person-job allocation processes.

John P. Campbell (1986) provided a general overview of the status of Project A at the American Psychological Association annual meeting in Washington, D.C. in August 1986. The research Campbell described was being accomplished by researchers at Human Resources Research Organization, American Institutes of Research, Personnel Decisions Research Institute, and the Army Research Institute.

The objectives of Project A given by Campbell were to:

1) Identify the constructs that constitute the universe of information available for selection/classification into entry level skilled jobs.

2) Develop a general model of performance for entry-level skilled jobs.

3) Investigate the construct validity of the "method" variance in job performance measures.

4) Describe the utility functions and the utility metrics that individuals actually use when estimating "utility of performance."

5) Estimate the degree of differential prediction across
   (a) major domains of predictor information (e.g., abilities, personality, interests),
   (b) major factors of job performance, and
   (c) different types of jobs.

6) Determine the extent of differential prediction across racial and gender groups for a systematic sample of individual differences, performance factors, and jobs.

7) Develop new statistical estimators of classification efficiency.

Nineteen MOS are being analyzed in depth, including nine MOS for which job specific hands-on, job-knowledge, and behaviorally-anchored rating performance measures were developed. The nine MOS are: administrative specialist, cannon crewman, infantryman, medical care specialist, military police, motor transport operator, radio teletype operator, tank crewman, and vehicle and generator mechanic.

Tasks included for performance measure development essentially consisted of refining existing Army task data defining job requirements described in the Army's Soldier's Manual, the Army Occupational Survey data and other references. Teams of subject matter experts from proponent schools along with scientific staff members successively narrowed several hundred tasks for each MOS to 30 tasks based on judgments of importance, clarity and difficulty. All 30 tasks for each MOS were incorporated into the job-knowledge tests; 15 of the tasks were considered appropriate for hands-on tests. The task selection procedure was well-defined and included decision rules, policy-capturing techniques and Delphi negotiations to assist in making the procedure systematic and as uniform as possible.

Table 25 lists the new predictors developed in Project A that were used for concurrent validation in 19 Army MOS. Table 26 lists the ability, temperament, and interest factors that served as predictors of job performance factors. The factor scores were simple sums of tests and inventory scales listed under each factor title.

If all rating scales and MOS-specific performance measures were aggregated at the task level and all major predictor sub-scores were used, about 200 criterion scores and 70 predictor scores would have been available for each individual. The decision to focus at the construct level coupled with budgetary considerations resulted both in a reduced predictor-criterion space for detailed analyses and the division of the sample into two groups or batches.

The 19 MOS of 400-600 individuals each were subdivided into two batches: Batch A (9 MOS) and Batch Z (10 MOS). The distinction between the two was that all of the criterion measures (aggregated at the construct level) were employed for each Batch A job. Performance measures used for Batch A only were: hands-on tests and paper-and-pencil job knowledge tests of MOS-specific task proficiency; rating scale measures of MOS-specific task proficiency that were also measured by hands-on and job-knowledge tests; and MOS-specific behaviorally-anchored rating scales representing major factors constituting job-specific technical and task proficiency. Performance measures common to Batch A and Batch Z were: behaviorally-anchored rating scales designed to measure non-job-specific performance; a rating scale of overall job-performance; a rating scale of noncommissioned officer potential; ratings of performance on representative "common" tasks; and paper-and-pencil tests of training achievement developed for each of the 19 MOS.

TABLE 25. SUMMARY OF NEW PREDICTOR MEASURES USED IN CONCURRENT VALIDATION SAMPLES FOR PROJECT A, ARMY RESEARCH INSTITUTE

COGNITIVE PAPER-AND-PENCIL TESTS

| Test Name (Construct represented) | Number of Items |
|---|---|
| Reasoning Test (Induction-figural reasoning) | 30 |
| Orientation Test (Spatial orientation) | 24 |
| Map Test (Spatial orientation) | 20 |
| Object Rotation Test (Spatial visualization - Rotation) | 90 |
| Assembling Objects Test (Spatial visualization - Rotation) | 32 |
| Maze Test (Spatial visualization - scanning) | 24 |

COMPUTER-ADMINISTERED TESTS

| Name (Construct represented) | Number of Items |
|---|---|
| Simple Reaction Time (Processing efficiency) | 15 |
| Choice Reaction Time (Processing efficiency) | 30 |
| Memory Test (Short-term memory) | 36 |
| Target Tracking Test #1 (Psychomotor precision) | 18 |
| Target Shoot Test (Psychomotor precision) | 30 |
| Perceptual Speed and Accuracy Test (Perceptual speed & accuracy) | 36 |
| Identification Test (Perceptual speed and accuracy) | 36 |
| Target Tracking Test #2 (Two-hand coordination) | 18 |
| Number Memory Test (Number operations) | 28 |
| Cannon Shoot Test (Movement judgment) | 36 |

(Continued)

TABLE 25. SUMMARY OF NEW PREDICTOR MEASURES USED IN CONCURRENT VALIDATION SAMPLES FOR PROJECT A, ARMY RESEARCH INSTITUTE (Continued)

NON-COGNITIVE PAPER-AND-PENCIL INVENTORIES

| Inventory Name and Subscale Name | Number of Items |
|---|---|
| Assessment of Background and Life Experiences (ABLE) Inventory | 209 |
| Adjustment | |
| Dependability | |
| Achievement | |
| Physical Condition | |
| Leadership | |
| Locus of Control | |
| Agreeableness/Likeability | |
| Army Vocational Interest Career Examination (AVOICE) | 176 |
| Realistic Interests | |
| Conventional Interests | |
| Social Interests | |
| Enterprising Interests | |
| Artistic Interests | |

Source: Campbell (1986).

Wise, Campbell, McHenry, and Hanson (1986) described, in a separate paper, a latent structure model of job performance factors. A set of 29 performance measures reflecting measurement methods and different aspects of job performance measures included hands-on performance measures on 15 tasks, five paper-and-pencil tests of job knowledge and school knowledge, supervisor and peer ratings of performance, and performance indicators contained in official personnel records and self-report questionnaires. A confirmatory factor analysis was used to determine a common set of performance measures for the nine jobs that accounted for individual differences on those measures.

TABLE 26. ABILITY, TEMPERAMENT, AND INTEREST FACTORS FOR PREDICTING JOB PERFORMANCE FACTORS PROJECT A, ARMY RESEARCH INSTITUTE

APTITUDE

- Technical Knowledge Factor
  ASVAB Auto/Shop Information
  ASVAB Mechanical Comprehension
  ASVAB Electronic Information

- Quantitative Skills Factor
  ASVAB Arithmetic Reasoning
  ASVAB Math Knowledge
  NEW Number Memory Accuracy

- Verbal Skill Factor
  ASVAB Word Knowledge
  ASVAB Paragraph Comprehension

- Cognitive Speed Factor
  ASVAB Coding Speed
  ASVAB Numerical Operations

- Spatial Ability Factor
  NEW Assembling Objects
  NEW Maze Test
  NEW Spatial Reasoning
  NEW Orientation Test
  NEW Map Test

INTERESTS

- Job Interest Factors
  NEW Combat Related
  NEW Technical/Professional
  NEW Construction
  NEW Food/Commissary
  NEW Audio-Visual Communication
  NEW Protective Service

(Continued)

TABLE 26. ABILITY, TEMPERAMENT, AND INTEREST FACTORS FOR PREDICTING JOB PERFORMANCE FACTORS PROJECT A, ARMY RESEARCH INSTITUTE (Continued)

MOTOR/PERCEPTUAL

- Psychomotor Ability Factor
  NEW Target Tracking 1
  NEW Target Tracking 2
  NEW Target School
  NEW Cannon Shoot

- Perceptual Speed-Accuracy Factor
  NEW Simple Reaction Time
  NEW Choice Reaction Time (Speed)
  NEW Target Identification (Speed)
  NEW Perceptual Speed
  NEW Short Term Memory (Speed)
  NEW Number Memory (Speed)
  NEW Target Identification (Accuracy)
  NEW Short Term Memory (Accuracy)
  NEW Number Memory (Accuracy)
  NEW Choice Reaction Time (Accuracy)

BIODATA/TEMPERAMENT

- Job Reward Preference Factor
  NEW Support for Individual
  NEW Routine
  NEW Autonomy

- Energy/Vitality Factor
  NEW Vitality (Energy/Dominance)

- Adjustment/Self-Control Factor
  NEW Adjustment (Stability/Self Esteem)
  NEW Socialization (Rule Abiding)

Source: Adapted from Campbell (1986).

Note. Factors identified from the analysis of concurrent validity on 9430 job incumbents.
Simple sum factor scores were formed from the tests and inventory scales listed under each factor title.

Table 27 summarizes the measurement methods and performance dimensions characterizing the common latent structure across the nine different jobs. The latent structure model specified included five job performance constructs shown in the table and two measurement method factors, a written test "method" factor, and a ratings "method" factor. The estimated mean of the intercorrelations among the construct scores was r = .40; the highest intercorrelation was r = .80 between technical job knowledge and general soldiering. The confirmatory analysis showed that the overall model fits extremely well.

The latent performance structure appears to be composed of very distinct performance components, and this suggests that different constructs would be predicted by different types of tests. Thus, validity levels may vary across performance constructs of a job. Validity levels for different predictor types may also vary across jobs that have radically different performance construct weights. There can be considerable variations in weights across job families and similar weights within a job family, providing the possibility of differential validity.

Wise et al. inferred that Leadership/Effort and Maintaining Personal Discipline dimensions are aspects of performance that are under motivational control and may be predicted best by personality and interest measures.

They concluded that the five-factor common latent structure is stable across jobs sampled from this population, and that in generalizing to a wider domain of jobs, it would be reasonable to suppose that different performance dimension structures might define different populations of jobs.

Table 28 defines the criterion measures that comprise each of the five performance dimensions used in the validation of existing and experimental tests. (See Appendix B for more detailed definitions of job performance dimensions.) The ASVAB measures were administered to subjects about two years before

TABLE 27. MEASUREMENT METHODS AND PERFORMANCE DIMENSIONS REPRESENTING THE COMMON LATENT STRUCTURE ACROSS ALL JOBS IN SAMPLE PROJECT A, ARMY RESEARCH INSTITUTE

| Measurement Methods | Performance Dimensions | Overall Performance |
|---|---|---|
| Hands-on MOS Specific Task Tests<br>Written MOS Specific Task Tests<br>Supervisor Ratings of Technical Skill | MOS-Specific Knowledge and Skill | JOB PERFORMANCE |
| Hands-On Tests of Common Soldier Tasks<br>Written Tests of Common Soldier Tasks | Basic Soldiering Knowledge and Skill | |
| Ratings of: Leadership/ Effort and Self-Development<br>Awards and Certificates<br>Combat Effectiveness Appraisals | Leadership and Effort | |
| Ratings of Discipline & Self-Control<br>Avoiding Article 15<br>Being Promoted On Time | Personal Discipline | |
| Ratings of Physical Fitness<br>Military Appearance<br>Physical Readiness Scores | Fitness Appearance | |

Source: Extracted from Wise et al. (1986).

TABLE 28. THE CRITERION MEASURES THAT COMPRISE EACH PERFORMANCE DIMENSION COMMON LATENT STRUCTURES ACROSS ALL JOBS IN SAMPLE, PROJECT A, ARMY RESEARCH INSTITUTE

1) Task Proficiency: MOS (Job) specific core technical skills:

The proficiency with which the individual performs the tasks which are "central" to his or her job (MOS). The tasks represent the core of the job and they are its primary definers from job to job.

The subscales representing core content in both the knowledge tests and the job sample tests that loaded on this factor were summed within method, standardized, and then added together for a total factor score. The factor score does not include any rating measures.

2) Task Proficiency: General or common skills:

In addition to the core technical content specific to an MOS, individuals in every MOS responsible for being able to perform a variety of general or common tasks--e.g., use of basic weapons, first aid, etc. This factor represents proficiency on these general tasks.

The same procedure (as for factor one) was used to sum the general task scales, standardized within methods, and add the two standardized scores.

3) Peer Leadership, Effort, and Self Development:

Reflects the degree to which the individual exerts effort over the full range of job tasks, perseveres under adverse or dangerous conditions, and demonstrates leadership and support toward peers. That is, can the individual be counted on to carry out assigned tasks, even under adverse conditions, to exercise good judgment, and to be generally dependable and proficient?

Five scales from the Army-wide BARS rating form (general technical performance, peer leadership, demonstrated effort, self development, general maintenance), the expected combat performance scales, the job specific BARS scales, and the total number of commendations and awards received by the individual were summed for this factor.

(Continued)

TABLE 28. THE CRITERION MEASURES THAT COMPRISE EACH PERFORMANCE DIMENSION COMMON LATENT STRUCTURES ACROSS ALL JOBS IN SAMPLE, PROJECT A, ARMY RESEARCH INSTITUTE (Continued)

4) Maintaining Personal Discipline:

Reflects the degree to which the individual adheres to Army regulations and traditions, exercises personal self control, demonstrates responsibility in day to day behavior, and does not create disciplinary problems.

> Scores on this factor are composed of three Army-wide BARS scales (adherence to traditions and regulations, exercising self control, demonstrating integrity, a subscale from the combat rating pertaining to avoidance of trouble, and two indices from the administrative records (number of disciplinary actions and promotion rate).

5) Military Bearing/Physical Fitness

Represents the degree to which the individual maintains an appropriate military appearance and bearing and stays in good physical condition.

> Factor scores are the sum of the physical fitness qualification score from the individual's personnel record and the "military bearing and appearance" rating scale.

Source: Campbell (1986)

the criterion measures were obtained. The experimental measures were obtained concurrently with the criterion measures.

In Table 29 the multiple validity correlations are shown for each of the five criterion factors averaged across the nine jobs. Several points are especially noteworthy. First, the validity of the ASVAB composite against the five criteria ranges from R = .67 to R = .19. The magnitude of the validity for MOS-Specific Technical Skills, R = .61, and for MOS-General Basic Soldiering skills, R = .67, are impressively high, possibly among the highest magnitude yet reported for ASVAB against job performance using a sizeable sample. ASVAB also shows quite useful multiple validities for the three other criterion factors, ranging from R = .19 to R = .35.

TABLE 29. MULTIPLE VALIDITY CORRELATION[a] OF FIVE INDEPENDENT PREDICTOR COMPOSITES WITH EACH OF FIVE JOB PERFORMANCE CRITERION FACTORS AVERAGED ACROSS NINE JOBS (N=4400)

| Predictor Composites | Job Performance Criterion Factors | | | | |
|---|---|---|---|---|---|
| | MOS Technical (Job Specific Core Skills) | Basic Soldiering (General Skills) | Leadership and Effort | Personal Discipline | Military Bearing and Physical Fitness |
| ASVAB[b] composite K=4 | .61 | .67 | .35 | .19 | .21 |
| Spatial abilities composite K=1 | .54 | .64 | .28 | .16 | .11 |
| Perceptually-psychomotor composite (computerized) K=5 | .49 | .56 | .27 | .14 | .11 |
| Temperament scales and biodata composite (ABLE) K=4 | .24 | .25 | .34 | .32 | .37 |
| Interest scales composite (AVOICE) K=6 | .33 | .37 | .26 | .15 | .12 |
| ASVAB composite + new predictors composite[c] | .64 | .70 | .45 | .37 | .42 |
| Validity gain of combination | .03 | .03 | .10 | .18 | .22 |

Source: Adapted from Campbell (1986).

[a] Multiple correlations adjusted for shrinkage and corrected for restriction in range.

[b] K= Number of predictor scores in the composite.

[c] Values obtained from ARI.

Validities for predictor composites cannot be accurately compared across factors unless validities have been corrected for criterion unreliabilities. Preliminary reliability estimates computed for each criterion factor score are: MOS Technical, r = .85; Basic Soldiering, r = .85; Leadership/Effort, r = .80; Personal Discipline, r = .80; and Military Bearing/ Physical Fitness, r = .80. The reliabilities for ratings are considerably higher in Project A than ordinarily reported in published studies because of averaging multiple raters and multiple rating scales for subcomponents comprising each criterion factor. All rating measures were obtained from about 2 supervisors and 3 peers for each ratee. Validities of the ASVAB composite, corrected for criterion unreliabilities, against the five criteria in the same order given for reliabilities are: R = .66; R = .73; R = .39; R = .21; and R = .23. While the magnitude of validities is slightly increased, the large validity variation across factors remain.

One concern with the validities reported for ASVAB is that they might be spuriously high for some of the criterion factors since both the ASVAB and job-knowledge tests are paper-and-pencil tests. Partialling out the paper-and-pencil methods factor reduces the multiple correlation from R = .61 to R = .45 for the MOS Technical Proficiency criterion and from R = .67 to R = .54 for the Basic Soldiering criterion. However, it is important to point out that, in this instance, the job knowledge tests were developed to be especially close-linked to the specific content and procedures of job tasks. Thus, the unadjusted multiple correlation coefficients may reflect more accurately the criterion space of interest. Also, the variance partialled out of the paper-and-pencil methods factor probably includes valid cognitive ability variance. A partial correlation coefficient removes the method variance from both the predictor and the criterion--the latter may be acceptable, but not the former.

Partialling the rating methods factor from the Leadership/Effort criterion and ASVAB raises ASVAB validity from R = .35 to R = .47. In this instance, the higher correlated residual criterion space would be less contaminated, but at the same time, partialling removes some valid variance.

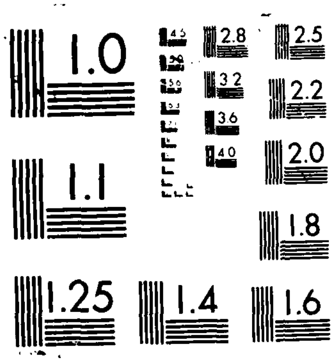
Second, the Leadership/Effort criterion factor is predicted well by a number of predictor types, with validities ranging from R = .35 to R = .26. The Personal Discipline and Military Bearing/Physical Fitness criterion factors are predicted well by the temperament and bio-data scales, R = .32 and R = .37.

Third, and of great significance, the gain in validity of the ASVAB composite achieved by combining other predictor composites with it adds from R = .03 to R = .22 correlational points. This finding provides strong empirical evidence of the value of combining other alternative predictors with an ability composite. The ASVAB composite combined with new composites shows a valiidity of R=.70 against basic soldiering and R = .45 against Leadership/Effort. These validity gains are especially significant for productivity increases in the Army's large manpower system.

Fourth, validity varies widely across factors, although every composite is valid for all factors. The ASVAB composite varies from R = .73 to R = .21 across the five criterion factors after correcting for unreliability. This finding seems to confirm the conjecture of Tenopyr and Oeltjan (1982):

> A matter which specifically should be a subject for future study is the effect of criteria upon validity generalization results. It appears that criteria such as supervisory ratings, which are subject to a large general factor, may lead to extensive generalizability, whereas those criteria which are more focused upon specific aspects of job behaviors and results may be associated with more situational specificity of validity, (p. 599).

AD-A186 602 THE VALIDITY OF SELECTION AND CLASSIFICATION PROCEDURES 2/2
FOR PREDICTING JOB PERFORMANCE(U) INSTITUTE FOR DEFENSE
ANALYSES ALEXANDRIA VA J ZEIDNER APR 87 IDA-P-1977
UNCLASSIFIED MDA903-84-C-0031 F/G 5/9 NL


1.0
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

Fifth, the pattern of validities obtained appear to indicate the possibility of obtaining differential validity or potential allocation efficiency (PAE) by the inclusion of new predictors in ASVAB and the expanded criterion space with different weights for each element by job family. More PAE may be derived from within both the cognitive domain and the biodata/ temperament domain by the development of PAE-oriented content or keys. (See discussion on PAE later in this report.)

A research question that still needs to be answered is how to determine the relative importance of each of the performance dimensions so that they can be combined into one overall measure of MOS performance. Such an overall measure will be used as the criterion against which predictors will be finally validated. Sadacca, deVerrra, DiFazo, and White (1986) reported on methodological considerations for weighting performance constructs. Trends indicate that the mean weights assigned to separate constructs by experts varied significantly in an MOS. Also, in considering the relative importance of the constructs across different MOS, the MOS-Specific Technical Skills construct received the highest weight and the Military Bearing/Physical Fitness construct received the lowest weight.

Jeffrey McHenry (1987) provided additional validity information on Project A predictor and criterion domains at the Society for Industrial and Organizational Psychology Second Annual Conference in Atlanta, Georgia in April 1987. Table 30 shows the composition of each predictor domain or composite. (It combines and aggregates the information given in Tables 25 and 26.) The technical, quantitative, verbal, and speed components of the ASVAB are generally recognized to constitute a very good measure of general cognitive ability (Hunter, et al., 1985). Five additional predictor domains are described in Table 30. These predictor domains were analyzed to determine the possibility of improving validities obtained by using the ASVAB composite alone.

TABLE 30. PROJECT A PREDICTOR DOMAIN, SOURCE AND COMPOSITE SCORES

| Predictor Domain | Source of Measure | Composite Score |
|---|---|---|
| General Cognitive Ability | Armed Services Vocational Aptitude Battery (ASVAB) | Technical<br>Quantitative<br>Verbal<br>Speed |
| Spatial Ability | Spatial Test Battery | Spatial |
| Perceptual-Psychomotor Ability | Computer Battery | Psychomotor<br>Complex Perceptual Speed<br>Complex Perceptual Accuracy<br>Number Speed and Accuracy<br>Simple Reaction Speed<br>Simple Reaction Accuracy |
| Temperament/ Personality | Assessment of Background and Life Experiences (ABLE) | Achievement Orientation<br>Dependability<br>Adjustment<br>Physical Condition |
| Vocational Interest | Army Vocational Interest Career Examination (AVOICE) | Skilled Technical<br>Structural/Machines<br>Combat-Related<br>Audiovisual Arts<br>Food Service<br>Protective Services |
| Job Reward | Job Orientation Blank (JOB) | Organizational and Co-Worker Support<br>Routine Work<br>Job Autonomy |

Table 31 shows the multiple validity correlations for each of the five criterion factors averaged across the nine jobs as does Table 29. Table 31, however, differs by reporting: validities for an auxiliary predictor composite, job reward preference (JOB); providing validities for an augmented perceptual-psychomotor composite (including 6 rather than 5 scores); and employing a sample size of N = 4039 (a more refined data set), rather than the larger sample size of N = 4400. Essentially the validities obtained are the same as in Table 29, varying within a few correlational points, except for a validity gain of .04 correlational points for the perceptual-psychomotor composite against Core Technical Proficiency and a loss of .04 correlational points for the ASVAB composite against Leadership/ Effort.

Overall the results in Table 31 show that the level of validities for MOS-Specific Technical Skills (R = .63) and for Basic Soldiering Skills (R = .65) criteria are as high as or higher than usually observed when ASVAB is correlated against training grades. Hunter et al. (1985) found a mean of r = .58 for 109 jobs across the four services. For the Leadership/ Effort criterion, the validity of the ASVAB composite is reasonably good (R = .31), as is the validity of the temperament/ biodata (ABLE) composite (R = .33). The validity of ABLE also is good for Personal Discipline (R = .32) and Military Bearing/ Physical Fitness (R = .37). The interest composite (AVOICE) predicts the first two task performance criteria (R = .35 and R = .34), but is not nearly as good for the last two non-task performance criteria, Personal Discipline (R = .13) and Military Bearing/Physical Fitness (R = .12). The pattern of composite validities for the third criterion, Leadership/Effort, appears to indicate that this criterion may have both proficiency and motivational components, with validities ranging for both cognitive and non-cognitive composites from R = .24 to R = .33, except for job reward preference, R = .19.

TABLE 31. MULTIPLE VALIDITY CORRELATION[a] OF SIX INDEPENDENT PREDICTOR COMPOSITES WITH EACH OF FIVE JOB PERFORMANCE CRITERION FACTORS AVERAGED ACROSS NINE JOBS (N=4039)

| Predictor Composites | Job Performance Criterion Factors | | | | |
|---|---|---|---|---|---|
| | MOS Technical (Job Specific Core Skills) | Basic Soldiering (General Skills) | Leadership and Effort | Personal Discipline | Military Bearing and Physical Fitness |
| ASVAB[b] composite K=4 | .63 | .65 | .31 | .16 | .20 |
| Spatial abilities composite K=1 | .56 | .63 | .25 | .12 | .10 |
| Perceptually-psychomotor composite (computerized) K=5 | .53 | .57 | .26 | .12 | .11 |
| Temperament scales and biodata composite (ABLE) K=4 | .26 | .25 | .33 | .32 | .37 |
| Interest scales composite (AVOICE) K=6 | .35 | .34 | .24 | .13 | .12 |
| Job reward preference (JOB) K=3 | .29 | .30 | .19 | .12 | .11 |

Source: Adapted from McHenry (1987).

[a] Multiple correlations adjusted for shrinkage and corrected for restriction in range.

[b] K= Number of predictor scores in the composite.

Table 32 examines the multiple validity correlations that result from combining the ASVAB composite (general cognitive ability), with the other types of predictor composites. The spatial ability composite adds several correlational points to the first two task performance-based criteria. However, the addition of the temperament/biodata composite (ABLE) to ASVAB is quite substantial: Leadership/Effort from R = .31 to R = .42; for Personal Discipline from R = .16 to R = .35; and for Military Bearing/Physical Fitness from R = .20 to R = .41. The other three predictor composites (perceptual-psychomotor, interest, and job reward preference) add little or no validity to the ASVAB composite across the five job factors.

Table 33 gives the multiple validity correlations that result from combining the experimental cognitive and non-cognitive components with the ASVAB composite. The results are similar to those obtained in Table 32, varying by one or two correlational points. Again, the results show that the ASVAB validities can only be augmented slightly by cognitive tests against the first two task performance-based criteria. ASVAB validities, however, can be augmented substantially by non-cognitive tests against the other three motivationally-based or partially motivationally-based criteria.

In Table 34 we see the impact of considering a composite of 24 cognitive and non-cognitive measures for each of the five job performance criteria. These results highlight the magnitude of validities obtainable by considering all test types together. The increments over ASVAB validities for the total combined composite of 24 cognitive and non-cognitive tests range from .04 correlational points for MOS-Specific Technical Skills to .22 correlational points for Military Bearing/Physical Fitness. Of equal interest is the actual level of validities reached against the five job performance criteria factors, with multiple correlations ranging from R = .37 to R = .70.

TABLE 32. MULTIPLE VALIDITY CORRELATION[a] OF COMBINED PREDICTOR COMPOSITES[b] WITH EACH OF FIVE JOB PERFORMANCE CRITERION FACTORS AVERAGED ACROSS NINE JOBS (N=4039)

| Predictor Composites | Job Performance Criterion Factors | | | | |
|---|---|---|---|---|---|
| | MOS Technical (Job Specific Core Skills) | Basic Soldiering (General Skills) | Leadership and Effort | Personal Discipline | Military Bearing and Physical Fitness |
| ASVAB[c] composite K=4 | .63 | .65 | .31 | .16 | .20 |
| ASVAB plus Spatial abilities composites K=5 | .65 | .68 | .32 | .17 | .22 |
| ASVAB plus Perceptually-psychomotor composites (computerized) K=10 | .64 | .67 | .32 | .17 | .22 |
| ASVAB plus Temperament scales and biodata composite (ABLE) K=8 | .63 | .66 | .42 | .35 | .41 |
| ASVAB plus Interest scales composite (AVOICE) K=10 | .64 | .66 | .35 | .19 | .24 |
| ASVAB plus Job reward preference (JOB) K=7 | .63 | .66 | .33 | .19 | .22 |

Source: Adapted from McHenry (1987).

[a] Multiple correlations adjusted for shrinkage and corrected for restriction in range.

[b] Combined predictor composites indicate the increases in $\underline{R}$ due to the addition of new predictor dimensions to the ASVAB general cognitive ability dimension.

[c] K= Number of predictor scores in the composites.

TABLE 33. MULTIPLE VALIDITY CORRELATION[a] OF COGNITIVE AND NON-COGNITIVE PREDICTOR COMPOSITES[b] WITH EACH OF FIVE JOB PERFORMANCE CRITERION FACTORS AVERAGED ACROSS NINE JOBS (N=4039)

| Predictor Composites | Job Performance Criterion Factors | | | | |
|---|---|---|---|---|---|
| | MOS Technical (Job Specific Core Skills) | Basic Soldiering (General Skills) | Leader-ship and Effort | Personal Discipline | Military Bearing and Physical Fitness |
| ASVAB[c] composite K=4 | .63 | .65 | .31 | .16 | .20 |
| New cognitive composites K=7 | .59 | .65 | .27 | .13 | .14 |
| ASVAB plus New cognitive composites K=11 | .65 | .69 | .32 | .17 | .23 |
| New non-cognitive composite K=13 | .44 | .44 | .38 | .35 | .38 |
| ASVAB plus New non-cognitive composite K=17 | .65 | .67 | .43 | .37 | .41 |

Source: Adapted from McHenry (1987).

[a] Multiple correlations adjusted for shrinkage and corrected for restriction in range.

[b] Combined cognitive predictor composites and combined non-cognitive predictor composites indicate the increases in $\underline{R}$ due to the addition of new predictor dimensions to the ASVAB general cognitive ability dimension.

[c] K = Number of predictor scores in the composite.

TABLE 34. MULTIPLE VALIDITY CORRELATION[a] OF COGNITIVE AND NON-COGNITIVE AND COMBINED PREDICTOR COMPOSITES[b] WITH EACH OF FIVE JOB PERFORMANCE CRITERION FACTORS AVERAGED ACROSS NINE JOBS (N=4039)

| Predictor Composites | Job Performance Criterion Factors | | | | |
|---|---|---|---|---|---|
| | MOS Technical (Job Specific Core Skills) | Basic Soldiering (General Skills) | Leader-ship and Effort | Personal Discipline | Military Bearing and Physical Fitness |
| Cognitive[b] composite K=11 | .65 | .69 | .32 | .17 | .23 |
| Non-cognitive composite K=13 | .44 | .44 | .38 | .35 | .38 |
| Combined cognitive and non-cognitive composites K=24 | .67 | .70 | .44 | .37 | .42 |

Source: Adapted from McHenry (1987).

[a] Multiple correlations adjusted for shrinkage and corrected for restriction in range.

[b] K= Number of predictor scores in the composite.

If performance is characterized as multidimensional as is characterized in Project A and a single, but different, criterion index is to be used for each job family, it still remains for the Army leadership to weigh the importance of each performance dimension. Research, along the lines of Sadacca, et al. (1986), can then specify the best method of combining the weighted factor dimensions into a composite reflecting the utility of total performance.

Some may argue that MOS-Specific Technical Skills and Basic Soldiering are by far the most important criterion factors in Project A and represent the real core of job performance. As noted earlier, these two criteria are correlated r = .80, the rank order of validities for them is identical, and their validity increments are smaller than for the other three factors. If these two factors receive high weights across all job families and also if, as argued by Hunter (in press), general cognitive abilities, rather than specific cognitive aptitudes, contribute most to such types of performance criteria, it may be difficult to achieve large increments to overall validity or to differential validity.

On the other hand, as Campbell (1986) indicates, Project A is guided by a view of job performance as being really multidimensional. He states that "There is not one outcome, one factor, or one anything that can be pointed to and labeled as job performance. It is manifested by a wide variety of behaviors, or things people do, that are judged to be important for accomplishing the goals of the organization." (p.7). Thus the concept of total performance is more than technical proficiency; it includes contributing to teamwork, continuing self-development, supporting the norms and customs of the organization and persevering in the fact of adversity.

The data already provided indicate that differential prediction occurs across the major components of performance being assessed in Project A. The data also indicate that there is differential prediction across the major components of the predictor universe in Project A. Further, the series of scaling studies conducted by Sadacca et al. (1986) show that judges can reliably indicate the relative importance of each criterion component within an MOS and that the patterns of weights seem to differ by MOS. Sadacca et al. are also scaling the utility of job performance by performance level combinations. To the extent such differential utilities can be measured, it potentially adds value to classification beyond differential prediction.

If tests are selected with PAE in mind as described in the section below, and given the multidimensional predictor and criterion space, and the differing utility of jobs by performance outcomes, a major research issue should be able to be resolved. In the words of Campbell "...one major research question we hope to answer is whether it is <u>ever</u> possible to estimate the parameters necessary for building a true classification algorithm. If it can't be done with a sample of 20 jobs and 500 cases per job then perhaps the textbook discussions of the classification problem are a bit academic." (p. 12).

Taken as a whole, the significance of this validation effort, when complete, should be great indeed. As the title of Campbell's 1986 paper suggests, it is even in this state of completion an example of results that can be achieved "When the Textbook Goes Operational." The results of Project A may become the standard for judging the operational effectiveness of all employment testing.

## M. POTENTIAL ALLOCATION EFFICIENCY

Brogden (1951) showed that differential assignment by means of a classification battery permits better utilization of personnel than does the use of a single weighted composite score. The use of a number of predictor composites from a classification battery identifies individuals with higher predictor scores because smaller selection ratios are called for than when a single composite is used, even when all personnel must be assigned. Brogden (1959) also showed that other factors constant, the utility gain from classification varies with the intercorrelations among estimates of job performance by the function $\sqrt{1-r}$. Even when the correlation among the estimates is high, considerable utility remains, e.g., when r = .80, classification gains are 45 percent as great as with intercorrelations of zero. Although not yet proven mathematically except for a number of simplifying assumptions by Brogden (1959), it has been shown through simulation studies that the higher the differential validity of each predictor composite, the higher the overall mean predicted performance when an optimal personnel assignment model is used (Niehl, 1967; Niehl and Sorenson, 1968; Sorenson, 1965; and Sorenson, 1967).

The utility of a classification battery, then, depends on the development of predictor composites such that each composite has relatively high validity for one job family and relatively low validities for other job families. PAE is a measure of the utility of composites derived from a classification battery. Small increments in PAE, measured in terms of average predicted performance can bring about worthwhile improvements in the military person-job match system.

Thus more findings concerning PAE, in the context of Project A's use of multidimensional criteria with differentially weighted components and use of predictors with heterogeneous test content, are required before a meaningful investigation of the utility of Project A products using alternative personnel allocation strategies can be completed.

There is a choice in the selection of tests for inclusion in the ASVAB: either validity generalization (a single general cognitive ability composite) or PAE can be maximized by the use of one of Horst's stepwise selection techniques against multiple criteria (here meaning across different jobs); either can be improved at the expense of the other in the initial selection of tests. However, the addition of tests with high PAE dces not need to detract from the validity generalization of a number of tests selected to maximize validity generalization, nor does the addition of tests with high validity generalization capability need to detract from PAE provided by tests already selected to maximize PAE.

Thus, the implementation of a selection/classification strategy that calls for selecting some tests to maximize the magnitude of validity coefficients and other tests to maximize PAE can achieve most of the PAE possible while losing little, if any, capability for validity generalization. Once a battery is selected the same weights are best for achieving either: the maximum average validity in accordance with validity generalization and without the use of a best allocation model; or the maximum average predicted performance across jobs using a best allocation model to take full advantage of PAE.

Future research should be undertaken (as it surely will be) to explore fully the very significant issue of how to best make use of whatever PAE exists through the application of assignment/ allocation models differing from the traditional linear program assignment models. The benefits obtainable from either the less understood multi-dimensional screening models when PAE exists, or the use of assignment models based on making optimal use of a single composite when PAE is non-existent should be fully explored.

It is commonly believed that the overall ability of a classification system is always maximized by maximizing the average validity across all jobs even where doing so reduces differential predictability. It is also believed that utility always is reduced by deliberately increasing PAE at the expense of average validity. Such beliefs are true only when the tests comprising the selection/classification battery have already been designated. Cecil Johnson, in a letter to me on this topic, elaborates:

> The utility of a classification battery is defined for the purposes of this discussion as being directly proportional to the average predicted performance of incumbents in a number of different jobs. When the test content of the selection/classification battery has been fully determined and only the selection of the test composites and weights for use in the selection and or classification of applicants for each job remains to be determined, the least square regression weights applied to all tests forming each test composite provides maximum utility. Such composites will not only provide the means of maximizing the average validities across jobs but will also maximize PAE. The validities of these composites are, of course, the multiple correlation coefficients between the battery and each job criterion measure. No set of composites selected to lower intercorrelations among composites nor to increase the variations of composite validities across jobs (as one might mistakenly attempt to do in order to increase PAE), can increase the utility function value. This statement is not true if the battery is in the process of being assembled and an optimal assignment model of the linear programming (LP) type is to be used to assign personnel to jobs from an already selected pool of personnel.
>
> The possibility of benefitting from a deliberate consideration of PAE, with some decrease in average validity as a consequence, depends to a large extent upon the following three conditions: (1) most importantly, whether the battery is fixed (already determined); (2) whether the selection/classification process is accomplished in one or two stages; and (3) which selection/classification model is being utilized to implement assignment to jobs.
>
> The latter two conditions can be delineated further:
>
> A Two stage model in which an assignment model allocates all personnel contained in a personnel pool to jobs; the pool may, or may not, have been produced by a single stage selection model.
>
> (1) LP type assignment model
> (2) Other type assignment model

B One Stage Models

(1) LP type assignment model, as in A(1) above, but with non-selection treated as if it were a dummy job; both selection and allocation are accomplished.

(2) Multidimensional screening model in which separate cut scores are set for each job category to produce the desired number of assignments to each job; personal preferences constrained by the cut scores effect both selection and allocation.

(3) Simple Selection Model.

I will first consider the two stage model as cited in A(1) above. It is relatively easy to demonstrate the superiority of the selection of tests for inclusion in a battery by explicit consideration of PAE, using a test selection process that trades-off magnitude of average validity to increase PAE. For example, in a hypothetical two test battery selected from a three test candidate set, the pair of tests with the greatest PAE can readily be shown to provide higher average predicted performance scores than another pair of tests that maximize average validity.

Harris (1967) conducted more comprehensive comparisons using batteries of tests selected from a set of 32 candidate tests. In Harris' study, model simulations were conducted on generated scores for hypothetical subjects based on sample characteristics of data obtained from 2480 actual subjects. For these actual subjects, both experimental test battery scores and training outcomes in 12 different courses were available.

Harris selected batteries of 5, 10, and 20 tests from the 32 tests for each of two alternative test selection methods. One of the two was Horst's "absolute" validity method which maximizes the sum of squares of the multiple correlation coefficients between the selected tests and each job criterion; the other was Horst's "differential" validity method which maximizes the sum of the multiple correlation coefficients between the selected tests and the differences among the job criteria.

Harris used separate sets of 10 samples of 216 entities (simulated people) in the investigation of each battery size. The same entities were used for both of the same sized batteries created with different selection approaches. The average predicted performance provided by the two alternative batteries when used in a personnel assignment model of the LP type showed about a 10 percent superiority of the "differential" selection method over the "absolute" method. The differences, for each of the three battery sizes, were significant at the .01 level. These results show the value of considering PAE at the expense of average validity in situation A(1).

The one stage model B(1) above has not been studied to the extent of A(1) model. I would expect an initial superiority of differential selection when no one is rejected. However, I would predict that the difference between the results of the two methods would decrease as the percentage placed in the not-select category goes up. This difference would pass through zero and finally end up in favor of Horst's absolute method as the percentage not-selected becomes large enough.

Much less is known about model B(2). However, I believe the consideration of PAE in selecting tests for inclusion in a battery for use in a multiple dimensional screening model will turn out to be important. The gain in utility from increasing PAE in a multiple screening model may come from the reduction of both recruiting costs and the costs of having to live with manpower shortages, although average predicted performance could also be increased by use of "differential" selection as compared to "absolute" selection.

(See Appendix C for procedures for determining PAE.)

# III. PREDICTIVE POWER OF COGNITIVE, NON-COGNITIVE AND ALTERNATIVE PREDICTORS

The proper estimates of the operational effectiveness or true validities of predictors in the employment setting should be based on corrections for (1) errors of measurement of the criteria and (2) restrictions in range of the predictors, if the validities are computed for incumbent groups. Failure to make both corrections, (especially in organizations such as the military that depend heavily on tests for the initial selection decision) will result in considerable underestimates of operational validities. In examining the major reviews described in this report, it was evident that only the validities of the U.S. Employment Service had been corrected for both errors in measuring the criteria and restriction in range (based on estimated values in a variety of earlier studies); military studies typically correct validities only for range restriction.

Fifteen years ago, Lent et al. (1971) reported that the median sample size in over 400 published validity studies was 68. Studies routinely show that 50 percent or more of the variance can be explained by sample size differences. Schmidt and Hunter (1981) state that about 85 percent of the artifactual variance in validities is accounted for by sampling error. Thus the problem of sample error in most studies of validity needs to be taken directly into account.

In order to obtain the kind of uniform data base needed to define more accurate validity levels for ability tests across job families, it is necessary to focus on cumulative studies of the military and Employment Service over the last four decades. In those studies, similar aptitude tests are used as predictors, similar clusters of jobs define a job family, fairly similar types of measures of training and job performance are used as criteria, and large numbers of independent validity coefficients using relatively large samples form the basis of each validity coefficient.

Hunter et al. (1985) examined the validity of various ASVAB composite scores and also derived a general cognitive ability score from ASVAB subtests that they found comparable to GATB cognitive ability on the basis of a confirmatory factor analysis. In an analysis of five large military studies that included 250 jobs, Hunter et al. found the validity of the GATB cognitive ability estimate was r = .55 and that the comparable military general cognitive ability was r = .60, a nine percent increase in validity. The higher validity of the ASVAB was attributed to better measures of verbal and quantitative aptitudes and the inclusion of technical aptitude (comprised of mechanical comprehension and electronics information tests). Hunter et al. state that the GATB validities are floor values for ASVAB validities. This is significant because it means that ASVAB has high validities for both civilian and military jobs.

The validities for cognitive abilities in the military and Employment Service sets of analyses are congruent, even when considering the broader sampling of job types included in the GATB studies.

With reference to alternative predictors, the Hunter and Hunter (1984) meta-analysis results provide the best overall estimates of the relative value of predictor types against supervisory ratings. Their analyses differentiate between predictors for entry level jobs and for promotion or certification. The four types of predictors that are best for entry level jobs are:

1. ability composite (R = .53)
2. job tryout (r = .44)
3. biodata (r = .37)
4. reference check (r = .26).

The conclusion, then, is that several types of alternative predictors show substantial validity against overall performance ratings for either entry levels jobs or for promotion/certification--but none has higher validity or as much practical utility as ability tests. A significant question that remains unanswered is how many correlational points can be added by combining non-cognitive with an ability composite.

Campbell (1986), McHenry (1987), and ARI provide data on combining non-cognitive predictors with ability tests. ASVAB validity increases from .03 to .22 correlational points against various criterion dimensions. Largest gains occur for those criteria that appear to be under motivational control, e.g., discipline, appearance and effort.

Table 35 summarizes the ASVAB and GATB validities for predicting training success and job performance across the total spectrum of jobs. These values emerge from synthesizing and judging the results reviewed in this report. The validity of a general cognitive ability composite for training is r = .65 and for job performance R = .53. When considering a combination of general cognitive ability and non-cognitive predictors, the validity increases to R = .67 against a criterion of specific-core technical skills.

It is worth noting that the validity of R = .67 is a mean across 9 different jobs; a mean of R = .57 is obtained with validities ranging from R = .47 to R = .83 across all the 19 jobs studied. The between-MOS validity differences for the same ASVAB predictor composites represent to some significant extent true differences in job requirements. These validities are the standard against which all predictors for entry level jobs may be compared.

Figure 1 shows some of the same information in graphic form. Over the years, employment tests have been frequently attacked on a variety of grounds, as in the assertion that they are poor predictors of job performance. Critics claimed that, at best, tests had "low or moderate validity" and that,

TABLE 35. SUMMARY OF ASVAB AND GATB AVERAGE VALIDITIES FOR TRAINING AND JOB PERFORMANCE

| | Validity | |
|---|---|---|
| Battery | Training Criteria | Job Performance Criteria |
| ACB Aptitude Areas (Army)[a] | .65 | -- |
| ASVAB Aptitude Area (Army)[b] | -- | .47 |
| ASVAB Aptitude Indices (Air Force)[c] | .65 | -- |
| ASVAB General Cognitive Ability (all military services)[d] | .60 | -- |
| ASVAB plus Non-Cognitive Predictors (Army)[e] | -- | .67 |
| GATB[f] | .57 | .53 |

[a] Maier and Fuchs (1972).

[b] McLaughlin et al. (1984).

[c] Weeks et al. (1975).

[d] Hunter et al. (1985).

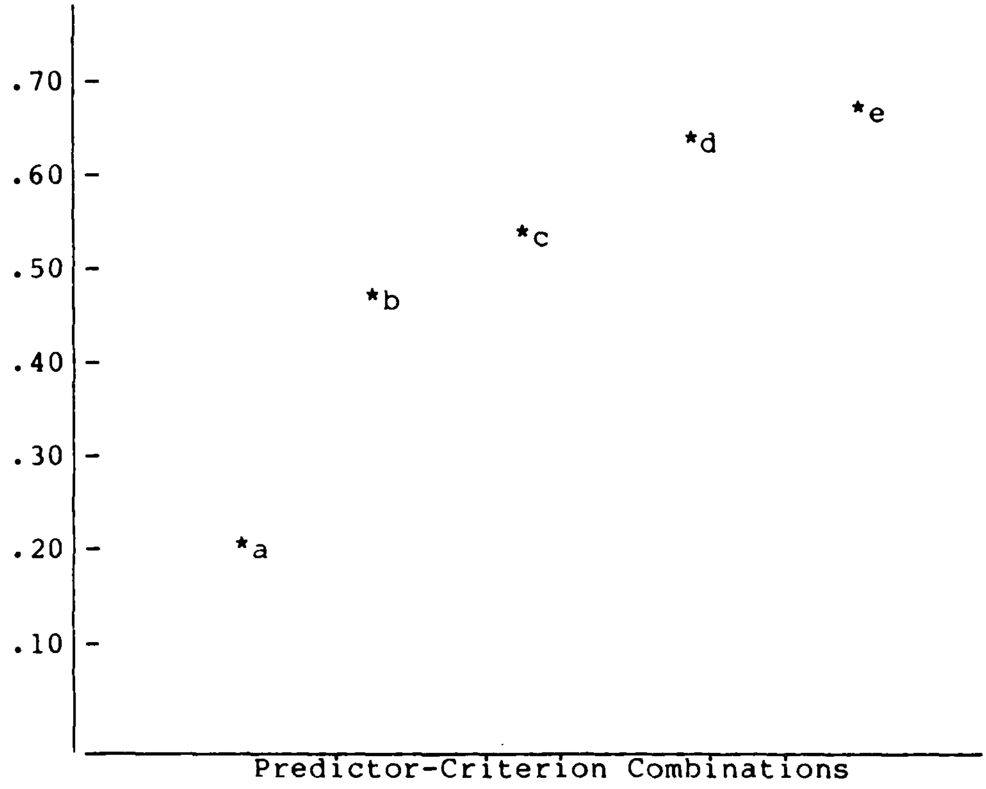
[e] McHenry (1987).

[f] Hunter (1983b).

in general, they "accounted for only five percent or so of job performance variance" (an indicator of utility not directly related to the benefit obtained in selection). And whatever the validity obtained, that value pertained only to a specific application.

Ghiselli's (1973) comprehensive findings, reporting a mean validity of r = .22, were typically cited as evidence of the low value of tests as predictors of job performance. Ghiselli

FIGURE 1. OBSERVED AND OPERATIONAL JOB VALIDITIES AS A FUNCTION OF PREDICTOR COMBINATIONS AND CRITERION TYPES



Note.
[a]Ghiselli (1973). Single ability, interest and personality tests; uncorrected; against ratings.
[b]McLaughlin et al. (1984). General cognitive ability composites; corrected for range restriction; against job knowledge.
[c]Hunter (1983b). Weighted general cognitive ability composite; corrected for range restriction and criterion unreliability; against ratings.
[d]McHenry (1987). Weighted general cognitive ability composite; corrected for range restriction; against job knowledge and hands-on performance composite criterion.
[e]McHenry (1987). Weighted composite of cognitive and non-cognitive predictors corrected for range restriction against job knowledge and hands-on performance composite criterion.

cautioned, however, that these were uncorrected validities for single tests of all types, and as observed validities they were underestimates of test effectiveness.

When a weighted composite of the same carefully dev loped and standardized general cognitive ability tests, e.g., GATB, corrected for range restriction and criterion unreliability is used as a predictor across a wide spectrum of jobs, the operational or true validity against rating criteria rises to R = .53.

When a weighted composite of cognitive ability tests and non-cognitive predictors is used against specially developed, high quality performance measures, e.g., reliable hands-on and job knowledge tests of specific technical and general proficiency, Project A ASVAB validity rises to R = .67. This level of validity, if accepted as the standard, is very important, since each percentage gain in validity represents a corresponding percentage increase in productivity, e.g., a validity gain of fourteen correlational points (or a 26 percent gain from R = .53 to R = .67) results in a 26 percent improvement in productivity.

# IV. THEORETICAL AND METHODOLOGICAL IMPLICATIONS

From the earliest days of personnel research, scientists sought to develop general principles of selection and testing. Many believed that only a cumulative approach would reveal the broad trends in the predictive power of tests. Throughout the years, collections of validity data showed wide variations. Despite his desire to find general traits, Ghiselli (1959) finally succumbed to the doctrine of situational specificity:

> A confirmed pessimist at best, even I was surprised at the variation in findings concerning a particular test applied to workers on a particular job. We certainly never expected the repetition of an investigation to give the same results as the original. But we never anticipated them to be worlds apart. Yet this appears to be the situation with test validities, (pp. 397-398).

Validity coefficients were believed to be specific to the situation in which they were determined and were not applicable to other situations, which could differ in location, time period, job content, organizational context, background variables, and the interaction of situational variables.

Schmidt and Hunter (1977, 1981) developed a Bayesian statistical model for testing the hypothesis that variations in validity coefficients in different studies were due to statistical artifacts. They found that most of the inconsistent findings across studies were the results of sampling error and failure to take into account other systematic effects such as error of measurement in criteria and predictors and restriction in range. A different view began to emerge--a view of validity generalization--validities could be extended to new situations.

Since the late 1970s, the validity generalization model has been applied to sets of validities in dozens of different occupations, thereby rejecting the concept that the validity of ability tests was job specific. These results have generally been well accepted by the scientific community. However, the older view of employment testing is so firmly entrenched in scientific thinking that questions continue to arise concerning the methodology of validity generalization, and the new results and conclusions that emerge from its application. Schmidt, Pearlman, Hunter, and Hirsh (1985) responded to these concerns in a 100 page question and answer debate in Personnel Psychology. Readers interested in knowing what the continuing technical and philosophical concerns are will find this article quite helpful.

## A. MAJOR CONCLUSIONS

This report reviewed the results of major validation and meta-analytic studies. The major conclusions reached are consistent with prevailing validity generalization statements, namely:

- Ability tests are valid for all jobs and job groupings.
- Against job performance criteria, the validity of cognitive ability decreases as job complexity decreases, while the validity of psychomotor ability increases with decreases in job complexity. Thus validity may be greatly improved by using combinations of cognitive and psychomotor abilities.
- Ability tests are valid for predicting both success in training and on-the-job.
- The validity of ability predictors can be demonstrated through validation at the job family level since validity changes very little with studies that differ in time, organizational setting, or small changes in job content.

- For entry level jobs, ability tests, in comparison with alternative predictors, are the best predictors of most types of job performance criteria conventionally used.
- Ability tests are also very close to being the best predictors of promotion and certification.
- One method of increasing overall validity across jobs is to combine non-cognitive predictors with an ability composite.

On the other hand, data indicate that the latent performance structure of a job consists of very distinct components. Thus it is reasonable to expect that different performance constructs (and measures of these constructs) would best be predicted by different domains of predictor information. Data indicate that validity levels vary widely across performance constructs within a job.

Standard 1.16 of the professional standards for psychological tests published by the American Psychological Association (APA, 1985) raises the issue of the extent to which validities can be generalized (transported) to a specific new situation. It states:

> When adequate local validation evidence is not available, criterion-related evidence of validity for a specified test use may be based on validity generalization from a set of prior studies, provided that the specified test-use situation can be considered to have been drawn from the same population of situations on which validity generalization was conducted. (Primary)
>
> Comment:
>
> Several methods of validity generalization and simultaneous estimation have proven useful. In all methods, the integrity of the inference depends on the degree of similarity between the local situation and the prior set of situations. Present and prior situations can be judged to be similar, for example, according to factors such as the characteristics of the people and job functions involved. Relational measures (correlations, regressions, success rates, etc.) should be carefully selected to be appropriate for the inference to be made, (pp. 16-17).

Validity generalizations probably can be made more confidently for ASVAB in the military context than in most other situations because of the vast number of criterion-related research findings accumulated for this battery.

The data needed to establish many predictor-criterion relationships are promising but sparse. Meta-analytic results can provide important guidelines for the direction of new research. One such research need is evaluating the same alternative predictors (cognitive and non-cognitive) together against the same reliable, relevant and comprehensive criteria, including such components as job knowledge, hands-on performance and supervisor/peer rating measures. As noted earlier, the Army Research Institute is currently conducting such a large-scale validation effort and has released some very promising preliminary findings.

The investigation of the benefits obtainable from a set of predictor variables possessing PAE with respect to job criteria consisting of differentially weighted criterion components has not, as yet, been reported. An investigation of whether such a finding can be sustained is essential to the determination of the operational utility of the Army Research Institute's large-scale validation effort.

# V. PRACTICAL IMPLICATIONS

Validity coefficients shown in major selection and classification programs are quite substantial and result in significant gains in workforce productivity as will be shown in a subsequent report on the economic benefits of predicting job performance. In large programs such as the military, where several hundred thousand individuals are selected and classified on the basis of tests, even very small increases in validity result in major productivity gains. Some practical and understandable means of transforming validity findings is needed so that better decisions can be made concerning the value of selection and classification strategies and policies.

The language of business, as has often been stated, is dollars and in recent years dozens of studies have been directed toward costing human resources, especially in the area of selection. Using validity selection procedures or differential predictors and rational allocation models results in significant dollar gains. For example, gains attributable to employing valid selection and/or classification devices for one year were: $350 thousand for 50 managers in one organization (Cascio and Silbey, 1979, and Cascio, 1982); $18 million for the Philadelphia Police Department (Hunter, 1980, and Schmidt and Hunter, 1981); $50 million for commissioned military officers and $442 million for enlisted personnel (Maier and Fuchs, 1973); potentially $8 billion for new federal employees (Schmidt, Hunter, Outerbridge and Trattner, 1986); and potentially $87.5 billion for the national economy as a whole (Hunter and Schmidt, 1982). Costing gains in productivity such as those noted here also is a subject to be considered in a subsequent report.

# VI. APPENDICES

## A. GLOSSARY

ability test  A test that measures the current performance or estimates future performance of a person in some defined domain of cognitive, psychomotor, or physical functioning.

achievement test  A test that measures the extent to which a person commands a certain body of information or possesses a certain skill, usually in a field where training or instruction has been received.

adaptive testing  A sequential form of testing in which successive items in the test are chosen based on the responses to previous items.

aptitude test  A test that estimates future performance on other tasks not necessarily having evident similarity to the test tasks. Aptitude tests are often aimed at indicating an individual's readiness to learn or to develop proficiency in some particular area if education or training is provided. Aptitude tests sometimes do not differ in form or substance from achievement tests, but may differ in use and interpretation.

assessment procedure  Any method used to measure characteristics of people, programs, or objects.

attenuation  The reduction of a correlation or regression coefficient from its theoretical true value due to the imperfect reliability of one or both measures entering into the relationship.

battery  A set of tests standardized on the same population, so that norm-referenced scores on the several tests can be compared or used in combination for decision making.

behavior[a]  Observable aspects of a person's activities.

classification  The act of determining which of several possible job assignments a person is to receive.

composite score A score that combines several scores by a specified formula.

concurrent criterion-related validity Evidence of criterion-related validity in which predictor and criterion information are obtained at approximately the same time.

construct A psychological characteristic (e.g., numerical ability, spatial ability, introversion, anxiety) considered to vary or differ across individuals. A construct (sometimes called a latent variable) is not directly observable; rather it is a theoretical concept derived from research and other experience that has been constructed to explain observable behavior patterns. When test scores are interpreted by using a construct, the scores are placed in a conceptual framework.

inter-rater reliability Consistency of judgments made about people or objects among raters or sets of raters.

interest inventory A set of questions or statements that is used to infer the interests, preferences, likes, and dislikes of a respondent.

inventory A questionnaire or checklist, usually in the form of a self-report, that elicits information about an individual. Inventories are not tests in the strict sense; they are most often concerned with personality characteristics, interests, attitudes, preferences, personal problems, motivation, and so forth.

item analysis The process of assessing certain characteristics of test items, usually the difficulty value, the discriminating power, and sometimes the correlation with an external criterion.

job analysis Any of several methods of identifying the tasks performed on a job or the knowledge, skills, and abilities required to perform that job.

job relatedness[a] The inference that scores on a selection instrument are relevant to performance or other behavior on the job; job relatedness may be demonstrated by appropriate criterion-related validity coefficients or by gathering evidence of the relevance of the content of the selection instrument, or of the construct measured.

linear combination[a] The sum of scores, whether weighted differentially or not, on different assessments to form a single composite score.

longitudinal study Research that involves the measurement of a single sample at several different points in time.

meta-analysis[a] A procedure to cumulate findings from a number of validity studies to estimate the validity of the procedure for the kinds of jobs or groups of jobs and settings included in the studies.

multivariate[a] Characterizing a measure or study that incorporates several variables.

norms Statistics or tabular data that summarize the test performance of specified groups, such as test takers of various ages or grades. Norms are often assumed to represent some larger population, such as test takers throughout the country.

norm-referenced test An instrument for which interpretation is based on the comparison of a test taker's performance to the performance of other people in a specified group.

objective[a] Pertaining to scores obtained in a way that minimizes bias or error due to different observers or scores.

percentile The score on a test below which a given percentage of scores fall.

performance[a] The effectiveness and value of work behavior and its outcomes.

personality inventory An inventory that measures one or more characteristics that are regarded generally as psychological attributes or interpersonal skills.

predictive criterion-related validity Evidence of criterion-related validity in which criterion scores are observed at a later date, for example, for job or school performance.

predictor A measurable characteristic that predicts criterion performance such as scores on a test, evidence of previous performance, and judgments of interviewers, panels, or raters.

projective technique A method of personality assessment in which the test taker provides free responses to a series of stimuli such as inkblots, pictures, or incomplete sentences. The term reflects the assumption that people project into their responses their perceptions, feelings, and styles. Also called projective method.

psychometric Pertaining to the measurement of psychological characteristics such as abilities, aptitudes, achievement, personality, traits, skill, and knowledge.

regression equation[a] An algebraic equation used to predict criterion performance from predictor scores.

relevance[a] The extent to which a criterion measure reflects important job performance dimensions or behaviors.

reliability The degree to which test scores are consistent, dependable, or repeatable, that is, the degree to which they are free of errors of measurement.

reliability coefficient A coefficient of correlation between two administrations of a test. The conditions of administration may involve variation of test forms, raters or scorers, or passage of time. These and other changes in conditions give rise to qualitying adjectives being used to describe the particular coefficient, e.g., parallel form reliability, rater reliability, test retest reliability, etc.

residual score The difference between the observed and the true or predicted score.

restriction of range A situation in which, because of sampling restrictions, the variability of data in the sample is less than the variability in the population of interest.

score Any specific number resulting from the assessment of an individual; a generic term applied for convenience to such diverse measures as test scores, estimates of latent variables, production counts, absence records, course grades, ratings, and so forth.

sample[a] The individuals who are actually tested from among those in the population to which the procedure is to be applied.

selection decision A decision to accept or reject applicants for a job on the basis of information.

selection instrument[a] Any method or device used to evaluate characteristics of persons as a basis for accepting or rejecting applicants.

selection procedures[a] Process of arriving at a selection decision.

shrinkage Refers to the fact that a prediction equation based on a first sample will tend not to fit a second so well.

shrinkage correction[a] Adjustment to the multiple correlation coefficient for the fact that the beta weights in a prediction equation cannot be expected to fit a second sample as well as the original.

skill[a] Competence to perform the work required by the job.

split-half reliability coefficient An internal analysis coefficient obtained by using half the items on the test to yield one score and the other half of the items to yield a second, independent score. The correlation between the scores on these two half-tests, stepped up via the Spearman-Brown Formula, provides an estimate of the alternate-form reliability of the total test.

standardized prediction[a] A test employed for estimating a criterion of job performance, the test having been developed and normative information produced according to professionally prescribed methods as described in standard reference works.

standard score A score that describes the location of a person's score within a set of scores in terms of its distance from the mean in standard deviation units.

test[a] A measure based on a sample of behavior.

test-retest coefficient A reliability coefficient obtained by administering the same test a second time to the same group after a time interval and correlating the two sets of scores.

unidimensionality A characteristic of a test that measures only one latent variable.

utility The relative value of an outcome with respect to a set of other possible outcomes.

validation The process of investigation by which the degree of validity of a proposed test interpretation can be evaluated.

validity The degree to which a certain inference from a test is appropriate or meaningful.

validity coefficient A coefficient of correlation that shows the strength of the relation between predictor and criterion.

validity generalization Applying validity evidence obtained *in one or more situations to other similar situations on* the basis of simultaneous estimation, meta-analysis, or synthetic validation arguments.

variability[a] The spread or scatter of scores.

variable A quantity that may take on any one of a specified set of values.

variance A measure of variability; the average squared deviation from the mean; the square of the standard deviation.

Z-score A type of standard score scale in which the mean equals zero and the standard deviation equals one unit for the group used in defining the scale.

---

Source: Adapted from AERA et al., Standards for Educational and Psychological Testing (1985), except for terms indicated by [a] adopted from SIOP, Principles for the Validation and Use of Personnel Selection Procedures (1987).

## B. DEFINITIONS OF THE JOB PERFORMANCE CONSTRUCTS

### Core Technical Proficiency

This performance construct represents the proficiency with which the soldier performs the tasks that are "central" to the job. The tasks represent the core of the job and they are the primary definers of the job. For example, the first tour Armor Crewman starts and stops the tank engines; prepares the loader's station; loads and unloads the main gun; boresights the M60A3; engages targets with the main gun; and performs misfire procedures. This performance construct does not include the individual's willingness to perform the task or the degree to which the individual can coordinate efforts with others. It refers to how well the individual can execute the core technical tasks the job requires, given a willingness to do so.

### General Soldiering Proficiency

In addition to the core technical content specific to a job, individuals in every job also are responsible for being able to perform a variety of general soldiering tasks (e.g., determines grid coordinates on military maps; puts on, wears, and removes M17 series protective masks with hood; determines a magnetic azimuth using a compass; collects/reports information--SALUTE; and recognizes and identifies friendly and threat aircraft). Performance on this construct represents overall proficiency on these general soldiering tasks. Again, it refers to how well the individual can execute general soldiering tasks, given a willingness to do so.

### Leadership, Effort and Self Development

This performance construct reflects the degree to which the individual exerts effort over the full range of job tasks, perseveres under adverse or dangerous condition3, and demonstrates leadership and support toward peers. That is, can the individual be counted on to carry out assigned tasks, even under adverse conditions, to exercise good judgment, and to be generally dependable and proficient? While appropriate knowledges and skills are necessary for successful performance, this construct is only meant to reflect the individual's willingness to do the job required and to be cooperative and supportive with other soldiers.

## Maintaining Personal Discipline

This performance construct reflects the degree to which the individual adheres to Army regulations and traditions, exercises personal self-control, demonstrates integrity in day-to-day behavior, and does not create disciplinary problems. People who rank high on this construct show a commitment to high standards of personal conduct.

## Military Bearing and Physical Fitness

This performance construct represents the degree to which the individual maintains an appropriate military appearance and bearing and stays in good physical condition.

## C. PROCEDURES FOR DETERMINING POTENTIAL ALLOCATION EFFICIENCY

The presence or absence of PAE in a test battery, for a particular set of jobs, can be determined by conducting a simulation study in which vectors of test scores are generated for each entity (simulated soldier). As many samples of entities as required by the experimental design can be readily created. Universe covariances among the tests and the validity of each test for each job must be estimated from empirical data and used as the basis for generating the score vectors for each entity. The sample covariances and validities then have the essential statistical properties of samples drawn from the universe defined from empirical data.

To complete the simulation, the predicted performance for each job is computed separately for each entity. The weights for use in computing test composite scores for each entity should be based on the universe data, a composite score computed for each entity/job and an optimal assignment algorithm used to assign each entity to a job--meeting sample quotas while maximizing the allocation sum (averaged predicted performance across all jobs in the sample).

If PAE is zero for a particular battery and set of jobs, the allocation sum across all samples of entities will not be different from the grand mean of all composite scores. Such a result would occur for a battery in which the best weighted composite (pertaining to each job using best unbiased weights) validities are not higher than the average validity of the composites that are best for other jobs. To the extent that existing aptitude composites used by the services closely approximate best weighted composites, the finding that validities associated with a job family are not higher than the validities of other aptitude composites for that family, one can be fairly certain that the application of the more rigorous test for PAE described above would also provide a finding of zero PAE for the battery and set of jobs.

## VII. REFERENCES

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education (1985). Standards for educational and psychological tests. Washington, D.C.: American Psychological Association.

Asher, J.J. & Sciarrino, J.A. (1974). Realistic work sample tests: A review. Personnel Psychology, 27, 519-533.

Boehm, V.R. (1982). Are we validating more but publishing less? (The impact of governmental regulation on published validation research--an explanatory investigation). Personnel Psychology, 35, 175-187.

Brogden, H.E. (1949). When testing pays off. Personnel Psychology, 2, 171-183.

Brogden, H.E. (1951). Increased efficiency of selection resulting from replacement of a single predictor with several differential predictors. Educational and Psychological Measurement, 11, 183-196.

Brogden, H.E. (1959). Efficiency of classification as a function of number of jobs, percent rejected, and the validity and intercorrelation of job performance estimates. Educational and Psychological Measurement, 19, 181-190.

Brogden, H.E. & Taylor, E.K. (1950). The dollar criterion: Applying the cost accounting concept to criterion construction. Personnel Psychology, 3, 133-154.

Campbell, J.P. (1986, August). Project A: When the textbook goes operational. Paper presented at The American Psychological Association Annual Meeting, Washington, D.C.

Cascio, W.F. (1982). Costing Human Resources: The Financial Impact of Behavior in Organizations. New York: Van Nostrand Reinhold.

Cascio, W.F. & Silbey (1979). Utility of the assessment center as a selection device. Journal of Applied Psychology, 64, 107-118.

Christal, R.E. (1976). What is the value of aptitude tests? Paper presented at the 18th Annual Conference of the Military Testing Association, Gulf Shores, AL.

Cohen, B., Moses, J.S. & Byham, W.C. (1974). The validity of assessment centers: A literature review. Pittsburgh, PA: Development Dimensions Press.

Committee on the Performance of Military Personnel Commission on Behavioral and Social Sciences and Education. National Research Council (1986). Wigdor, A.K. & Green, B.F. (Eds.). Assessing the Performance of Enlisted Personnel. Evaluation of a Joint-Service Research Project. Washington, D.C.: National Academy Press.

Dunnette, M.D. (1972). Validity study results for jobs relevant to the petroleum refining industry. Washington, D.C.: American Petroleum Institute.

Dunnette, M.D. & Borman, W.C. (1979). Personnel selection and classification. Annual Review of Psychology, 30, 477-525.

Eaton, N.K., Hanser. L.M. & Shields, J.L. (1986). Validating selection tests against job performance. In J. Zeidner (Ed.) Human productivity enhancement: Organizations, personnel and decision making. (pp. 382-438). New York: Praeger.

Fine, S.A. (1955). A structure of worker functions. Personnel and Guidance Journal, 34, 66-73.

Ghiselli, E.E. (1959). The generalization of validity. Personnel Psychology, 12, 397-402.

Ghiselli, E.E. (1966). The validity of occupational aptitude tests. New York: Wiley.

Ghiselli, E.E. (1973). The validity of aptitude tests in personnel selection. Personnel Psychology, 26, 461-477.

Haney, W. (1981). Validity, vaudeville and values. A short history of social concerns over standardized testing. American Psychologist, 36, 1021-1034.

Harris, R.N. (1967, March). A model sampling experiment to evaluate two methods of test selection. (Research Memorandum, 67-2), Statistical Research and Analysis Division, Washington, D.C.: U.S. Army Behavior and Systems Research Laboratory.

Hunter, J.E. (1980a). *Test validation for 12,000 jobs. An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB).* Washington, D.C.: U.S. Employment Service. U.S. Department of Labor.

Hunter, J.E. (1980b). *An analysis of validity, differential validity, test fairness, and utility for the Philadelphia police officer selection examination prepared by Educational Testing Service.* Unpublished manuscript, Michigan State University.

Hunter, J.E. (1981, April). *False premises underlying the 1978 uniform guidelines on employee selection procedures: The myth of test invalidity.* Paper presented to the Personnel Testing Council of Metropolitan Washington. Washington, D.C.

Hunter, J.E. (1983a). A causal analysis of cognitive ability, job knowledge, job performance and supervisor ratings. In F. Landy, S. Zedeck, J. Cleveland (Eds.). *Performance Measurement and Theory.* Hillsdale, NJ: Lawrence Earlbaum and Associates.

Hunter, J.E. (1983b). *Test validation for 12,000 jobs: An application of job classification and validity generalization analysis to the General Aptitude Test Battery (GATB).* Washington, D.C.: Division of Counseling and Test Development, Employment and Training Administration, U.S. Department of Labor.

Hunter, J.E. (in press). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior.*

Hunter, J.E., Crosson, J.J., & Friedman, D.H. (1985). *The validity of the Armed Services Vocational Aptitude Battery for civilian and military job performance.* Rockville, MD: Research Applications, Inc.

Hunter, J.E. & Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*, 72-98.

Hunter, J.E. & Schmidt, F.L. (1982). Fitting people to jobs: The impact of personnel selection on national productivity. In E.A. Fleishman and M.D. Dunnette (Eds.), *Human performance and productivity: Vol. I: Human capability assessment.* (pp. 232-284). Hillsdale, NJ: Erlbaum.

Jacobs, R., Kafry, D., & Zedeck, S. (1980). Expectations of behaviorally anchored rating scales. Personnel Psychology, 33, 595-640.

Kane, J.S. & Lawler, E.E. (1978). Methods of peer assessment. Psychological Bulletin, 85, 555-586.

King, L.M., Hunter, J.E., & Schmidt, F.L. (1980). Halo in a multidimensional forced-choice performance evaluation scale. Journal of Applied Psychology, 65, 507-516.

Kyllonen, P.C. (1986, January). Theory-based cognitive assessment. (AFHRL-TP-85-30). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Lent, R.H., Aurbach, H.A., & Levin, L.S. (1971). Research design and validity assessment. Personnel Psychology, 24, 247-274.

Lilienthal, R.A. & Pearlman, K. (1983). The validity of federal selection tests for aid technicians in the health, science, and engineering fields. (OPRD 83-1). Washington, D.C.: U.S. Office of Personnel Management, Office of Personnel Research and Development. (NTIS No. PB83-202051).

Maier, M.H. & Fuchs, E.F. (1972, September). Development and evaluation of a new ACB and aptitude area system. Technical Research Note 239. Military Selection Research Division, Washington, D.C.: U.S. Army Behavior and Systems Research Laboratory.

Maier, M.H. & Fuchs, E.F. (1973, September). Effectiveness of selection and classification testing. Research Report 1179. Individual Training and Manpower Development Technical Area, Alexandria, VA: U.S. Army Research Institute.

Maier, M.H. & Grafton, F.C. (1981, May). Aptitude composites for ASVAB 8, 9 and 10. Research Report 1308. Personnel Utilization Technical Area, Alexandria, VA: U.S. Army Research Institute.

Maier, M.H. & Hiatt, C.M. (1984, May). An evaluation of using job performance tests to validate ASVAB qualification standards. (CNR 89). Alexandria, VA: Center for Naval Analyses.

McDaniel, M.A. Schmidt, F.L., Raju, N.S., and Hunter, J.E. (1986). Interpreting the results of meta-analytic research: a comment on Schmitt, Gooding, Noe, and Kirsch (1984). Personnel Psychology, 39, 141-148.

McHenry, J.J. (1987, April). Project A validity results: The relationship between predictor and criterion domains. Paper presented at the Society for Industrial and Organizational Psychology Annual Conference. Atlanta, GA.

McLaughlin, D.H., Rossmeissl, P.G., Wise, L.L., Brant, D.A., & Wang, M.M. (1984, October). Validation of current and alternative ASVAB area composites, based on training and SQT information of FY81 and FY82 enlisted accessions. (Technical Report 651), Alexandria, VA: U.S. Army Research Institute.

Niehl, E. (1967, February). A general computer simulation for conducting allocation experiments. Research Memorandum 67-1. Washington, D.C.: U.S. Army Research Office. (AD A079 330).

Niehl, E. & Sorenson, R.C. (1968, January). Simpo-I entity model for determining the qualitative impact on personnel policies. Washington, D.C.: Technical Research Note 193. Statistical Research and Analysis Division, U.S. Army Behavior and Systems Research Laboratory. (AD 831 268).

O'Leary, B.S. (1980). College grade point average as an indicator of occupational success: An update. (PRR-80-23). Washington, D.C.: U.S. Office of Personnel Management, Personnel Research and Development Center. (NTIS No. PB81-121329).

Pearlman, K. (1982). The Bayesian approach to validity generalization: A systematic examination of the robustness of procedures and conclusions. (Doctoral dissertation, George Washington University.) Dissertation Abstracts International, 4960-B.

Pearlman, K., Schmidt, F.L., & Hunter, J.E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. Journal of Applied Psychology, 65, 373-406.

Personnel Research Section, Classification and Replacement Branch, The Adjutant General's Office. (1945). The Army General Classification Test. Psychological Bulletin, 42, 760-768.

Reilly, R.R. & Chao, G.T. (1982). Validity and fairness of some alternative employee selection procedures. Personnel Psychology, 35, 1-62.

Sadacca, R., deVera, M.V., DiFazio, A.S., White, L.A. Weighting performance constructs in composite measures of job performance. (1986, August). Paper presented at The American Psychological Association Annual Meeting, Washington, D.C.

Schmidt, F.L., Gast-Rosenberg, I., & Hunter, J.E. (1980). Validity generalization results for computer programmers. Journal of Applied Psychology, 65, 643-661.

Schmidt, F.L. & Hunter, J.E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Schmidt, F.L. & Hunter, J.E. (1981). Employment testing: Old theories and new research findings. American Psychologist, 36, 1128-1137.

Schmidt, F.L., Hunter, J.E., & Caplan, J.R. (1981). Validity generalization results for two job groups in the petroleum industry. Journal of Applied Psychology, 66, 261-273.

Schmidt, F.L., Hunter, J.E., McKenzie, R.C. & Muldrow, T.W. (1979). Impact of valid selection procedure on work force productivity. Journal of Applied Psychology, 64, 609-626.

Schmidt, F.L., Hunter, J.E., & Outerbridge, A.N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. Journal of Applied Psychology, 71, 432-439.

Schmidt, F.L., Hunter, J.E., Outerbridge, A.N., & Trattner, M.H. (1986). The economic impact of job selection methods on size, productivity, and payroll costs of the federal work force: an empirically based demonstration. Personnel Psychology, 39, 1-39.

Schmidt, F.L., Hunter, J.E., & Pearlman, K. (1982). Assessing the economic impact of personnel programs on work force productivity. Personnel Psychology, 35, 333-347.

Schmidt, F.L., Pearlman, K., Hunter, J.E. & Hirsh, H.R. (1985). Forty questions and validity generalizations and meta-analysis. Personnel Psychology, 38, 697-798.

Schmitt, N., Gooding, R.Z., Noe, R.D., Kirsch, M. (1984). Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics Personnel Psychology, 37, 407-422.

Schmitt, N. & Schneider, B. (1983). Current issues in personnel selection. *Personnel and Human Resources Management*, *1*, 85-125.

Schwab, D.P., Heneman, H.G., & De Cotiis, T.A. (1975). Behaviorally anchored rating scales: A review of the literature. *Personnel Psychology*, *28*, 549-562.

Society for Industrial and Organizational Psychology (1987). *Principles for the validation and use of personnel selection procedures*. College Park, MD: Society for Industrial and Organizational Psychology.

Sorenson, R.C. (1965, November). *Optimal allocation of enlisted men--full regression equations vs aptitude area scores.* Technical Research Note 163. Washington, D.C.: U.S. Army Personnel Research Office. (AD 625 224)

Sorenson, R.C. (1967, February). *Amount of assignment information and expected performance of military personnel.* Technical Research Report 1152. Washington, D.C.: U.S. Army Personnel Research Office. (AD 649 907)

Tenopyr, M.L. & Oeltjen, P.D. (1982). Personnel selection and classification. *Annual Review of Psychology*, *33*, 581-618.

Vineberg, R., & Joyner, J.N. (1982). *Prediction of job performance: Review of military studies*. Alexandria, VA: Human Resources Research Organization.

Weeks, J.L., Mullins, C.J., & Vitola, B.M. (1975, December). *Airman classification batteries from 1948 to 1975: A review and evaluation* (AFHRL-RT-75-78). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory. AD-A026 470

Wise, L.L., Campbell, J.P., McHenry, J.J. & Hansen, L.R. (1986, August). *A latent structure model of job performance factors*. Paper presented at The American Psychological Association Annual Meeting, Washington, D.C.

END

JAN.

1988

DTIC